Investigation of genomic and phenotypic plasticity in the lignocellulose-bioconverting and xylose-fermenting yeast *Scheffersomyces stipitis*

Samuel Vega Estévez

2020

A thesis submitted to the University of Kent for the Degree of Doctor of Philosophy in Genetics.

School of Biosciences

Faculty of Sciences

University of Kent

Declaration

No part of this thesis has been submitted in support of an application for any degree or other qualification of the University of Kent, or any other University or Institution of learning.

Samuel Vega Estévez.

September 2020.

To my dad, who will always be with me.

To my mum and my brother.

A mi padre, que estará siempre conmigo.

A mi madre y mi hermano.

“Evolution forged the entirety of sentient life on this planet using only one tool: the mistake.”

Dr Robert Ford.

Westworld, S1 e01. The original

ACKNOWLEDGEMENTS

Well… it seems that this is coming to an end. Time flies. I think that these three years have been the fastest/slowest ones that I can remember… Time does fly.

Any person that knows me at least a little bit realises that I am not characterised by a great ability to expose my emotions, and yet I think that if I am ever going to make an exception, this would not be a bad moment. So, Spoiler alert! This might be emotional?

The first people I would like to thank are my parents. Because they have always been there for me, supporting me even when I was not making it easy, and whose capacity to overcome difficulties and wish to give their children as much as they could have always been an example for me, and is something I have never said to them. To my brother David, for being the best older brother I could have (60% of the time), and my sister-in-law, Samara who has become an importance piece of my family. To my aunt Loli and my uncle Lito.

I also want to thank my supervisor, Dr. Alessia Buscaino, not just for giving me the chance to work in this (cool) project and for her supervision, but for being there for me in the last three years, helping me to become a better scientist and giving advice and support when I needed it. It has definitively been a pleasure working with you (most of the time).

The rest of the members in the AB Lab have been also very important, Misty and Jordan, who were there when I started and that were very helpful in my beginnings, and Marzia and Enea for the great moments in the lab. We form a great Mediterranean Famiglia. I would like to include here the rest of the member of the KFG, because despite I complain a lot (about them, to them) we are always there for each other, forming a small, weird, great, family. I would also like to thank Andrew Armitage for his help (and patience) in the bioinformatic analysis of this project.

Obviously, my friends have been a fundamental part in me finishing this project without completely losing my sanity. Marzia, for being a great friend and housemate (special thanks for not killing me this last year). The group “Dog, Lunch, Walking and GOOSE”, (quite self-explanatory): Emma, Kevin, Ruk and Sarah. Ane and Antoine for the nights of chats and alcohol “to forget”. Vanessa for the supporting messages, the bioinformatic advice, and for all the dog memes in which our conversations can be based during days. Timothy and Victor for their support, and Pablo for “obliging” me to go for walks and for giving me food to keep me in a good mood. Shamira, Raquel, Raquel and Samu for all these years of friendship.

This might sound a bit cheesy and unconventional, but thanks to all the people of my country that fought for the right of my generation to access a public, high quality High Education system, because without them many of us might not be able to work in what we want. Also, to all those who are still fighting to maintain it.

Finally, I would like to thank the University of Kent for their funding, because without it my PhD would have not been possible.

AGRADECIMIENTOS

Parece que tres años llegan a su fin. El tiempo vuela. Creo que puedo decir que estos tres años han sido los más rápidos/lentos que recuerdo… Sí, el tiempo vuela.

Cualquier persona que me conozca sabe que no me caracterizo por tener una gran capacidad de expresión en lo que a lenguaje emocional se refiere. Sin embargo, si alguna vez tenía que hacer una excepción, supongo que este no es un mal momento.

Las primeras personas a las que quiero agradecer son mis padres. Porque siempre me han apoyado, sin importar lo difícil que fuera y a pesar de lo mucho que yo en ocasiones lo dificultara. Su esfuerzo de superación y su deseo de dar a sus hijos todo lo que estaba en sus manos ha sido siempre un ejemplo para mí, y es algo que nunca les he dicho. A mi hermano David, que ha sido el mejor hermano mayor que podía tener (el 60% de las veces), y a mi cuñada Samara, que se ha convertido en una hermana para mí. A mis tíos Loli y Lito, por quererme tanto.

También quiero agradecer a mi supervisora, la Dra. Alessia Buscaino, no sólo por su haberme dado la oportunidad de hacer este proyecto y por su supervisión los últimos años, sino por de verdad haber estado a mi lado, enseñándome a ser un mejor científico, aconsejándome y apoyándome cuando lo necesitaba, sobre todo en los momentos duros. Ha sido un placer trabajar contigo.

Han sido también muy importantes para mí el resto de miembros del AB Lab, tanto los que estaban en mis comienzos, Misty y Jordan, que fueron de gran ayuda durante mi aprendizaje. También Marzia and Enea, por los momentos de ánimo y diversión. Juntos nos hemos convertido en la pequeña Famiglia mediterránea. Quiero incluir aquí al resto de miembros del KFG, porque a pesar de que me quejo mucho (de ellos, a ellos) siempre estamos ahí para apoyarnos unos a otros. También quisiera agradecer a Andrew Armitage por la ayuda prestada (y por tener paciencia conmigo).

Lógicamente, mis amigos han sido una parte indispensable en que yo pudiera acabar este proyecto sin haber perdido completamente la cordura. Marzia, por ser una gran amiga y compañera de piso (y no haberme matado durante este último año). El grupo “Dog, Lunch, Walking and GOOSE” (Emma, Sarah, Kevin and Ruk) por los momentos juntos. Ane and Antoine por las noches de charla y beber para olvidar. Vanessa por los mensajes de ánimos en el último año, los consejos bioinformáticos, y sobre todo los memes de perritos, en los que se pueden basar varios días de conversación. A Timothy por los mensajes de ánimo, Victor por los cafés y su gran apoyo, y Pablo por “obligarme” a dar paseos y darme comida para mantenerme de buen humor. Shamira, Raquel, Raquel y Samu por tantos años de amistad.

Quizá algo cursi y poco convencional, pero me gustaría agradecer a la gente de mi país que luchó por el derecho a que mi generación pudiera acceder a una universidad pública y de calidad, porque sin ellos muchos no habríamos podido tener el privilegio de dedicarnos a nuestra pasión.

TABLE OF CONTENTS

ABSTRACT

Eukaryotic genomes are often described as stable structures with well-preserved chromosome organisation. Consequently, genome instability is viewed as a deleterious event. However, it is becoming increasingly clear that genomes can be plastic, and that genome instability can be advantageous for adaptation to challenging environment.

The overall goal of this project was to understand whether and how genome plasticity contributes to environment adaptation in *Scheffersomyces stipitis,* one of the most promising yeast for the production of second-generation bioethanol. *S. stipitis* belongs to the CTG clade, formed several yeasts (such as *C. albicans*) whose genomes have been widely described as plastic.

To do so, the karyotype of 27 strains isolated from different habitats and countries was studied by CHEF electrophoresis. This allowed us to discover that two of the most used strains in *S. stipitis* research (NRRL Y-11545 and NRRL Y-7124) exhibited differences in their genome structures. Therefore, we compared the publicly available genome of the strain Y-11545 to the genome sequence of the natural isolate Y-7124, sequenced in this study by a hybrid TGS approach, based on the combination of Illumina and Oxford Nanopore Technologies (ONT). Moreover, to elucidate whether genomic plasticity conditions short-term evolution in *S. stipitis,* we analysed the karyotype of strains derived from the natural isolate Y-7124 after an adaptation experiment.

Our results prove that the genome of *S. stipitis* is plastic, since different genome conformations are observed among the natural isolates studied. We have also shown that this plasticity is strongly associated to repetitive regions, at least in the two natural isolates whose genome is available. The major contributors to repetitive regions in *S. stipitis* genome are transposable elements, altogether with subtelomeres, telomeres, (all of them described in this study for the first time) and centromeres. Moreover we have detected several chromosome modifications: (i) A reciprocal chromosome translocation between the natural isolates Y-11545 and Y-7124 (ii) An aneuploidy consisting of an extra chromosome in the evolved strain Y-50859, (iii) A reciprocal chromosome translocation between the strain Y-7124 and its derived strain Y-50859. Both modifications (i) and (ii) are associated to repetitive DNA.

In conclusion this project has demonstrated that the genome of *S. stipitis* is highly variable, and this plasticity is conditioned by repetitive DNA. Moreover, we have shown that strains with improved phenotypic traits related to fermentation are influenced by genome conformational changes.

LIST OF ABBREVIATIONS

ADPr: Adenosine diphosphate (ADP) ribose

AFEX CSH: Ammonia fibre expansion-pretreated corn stover hydrolysate

ALE: Adaptive laboratory evolution

AMP: Adenosine monophosphate

AP: Aspartic protease

BCATs: Branch chain aminotransferases

BER: Base excision repair

BIR: Break-induced replication

BUSCO: Benchmarking Universal Single Copy Orthologs

C5: 5-carbon sugars

C6: 6-carbon sugars

CCNV: Chromosomal copy number variation

CDEs: Centromeric DNA elements

CEN: Centromere

aCGH: Array comparative genomic hybridization

CHEF: Contour-clamped homogeneous electric field

CNVs: Copy number variants

Cox: Cytochrome c oxidase

DNA: Deoxyribonucleic acid

cDNA: Complementary DNA

ssDNA: single stranded DNA

DSB: Double strand break

FISH: Fluorescence in situ hybridisation

GCRs: Gross chromosomal rearrangements

GQ: G-quadruplexes

GWAS: Genome-wide association studies

HJ: Holliday junction

HR: Homologous recombination

INT: Integrase

JGI: Joint Genome Institute

KOG: Eukaryotic orthologous groups

LAP: LTR-associated protein

LARD: Large retrotransposon derivative

LINEs: Long interspersed elements

LTR: Long terminal repeats

LOH: Loss of heterozygosity

MAT: Mating-type

MFS: Major facilitator superfamily

MITEs: Miniature inverted repeat transponsable elements

MMR: Mismatch repair

NAHR: Non allelic homologous recombination

NER: Nucleotide excision repair

NGS: Next generation sequencing

NHEJ: Non-homologous end joining

NRRL: Norther regional research laboratory

OD: Optical density

ODM: Optimum defined media

ORF: Open reading frame

PhIGs: Phylogenetically inferred group comparison

PLEs: *Penelope*-like elements

PPP: Pentose phosphate pathway

qPCR: quantitative polymerase chain reaction

PSGHL: Dilute acid-pretreated switchgrass hydrolysate liquor

QTL: Quantitative trait loci

RF: Replication fork

RH: RNase H

RNA: Ribonucleic acids

RT: Reverse transcriptase

RTE: RAD52-dependent recombinational telomere elongation

SC: Synthetic complete

SGH: Switchgrass hydrolysate

SHAM: Salicylhydroxamic acid

SINEs: Short interspersed elements

SNP: Single nucleotide polymorphism

SOLiD: Sequencing by oligonucleotide ligation and detection

SRD: Split direct repeats

Sto: Alternative terminal oxidase

TAM: Transcription associated mutations

TAR: Transcription-associated recombination

TCA: Tricarboxylic acid

TE: Transposable element

TERC: Telomerase RNA component

TERT: Telomerase reverse transcriptase

TERRA: Telomeric repeat containing RNA

TGS: Third generation sequencing

TIRs: Terminal inverted repeats

Tps5: Ty5-like LTR retrotransposons in *S. stipitis*

TSD: Target site duplication

VIL: Valine, Isoleucine and Leucine

WGD: Whole genome duplication

XDH: Xylitol dehydrogenases

XK: Xylulokinases

XR: Xylose reductases

YPD: Yeast Extract-Peptone-D-Glucose

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

1.1 FROM GENES TO GENOME: YEAST AS A MODEL OF STUDY

One of the most fascinating properties of living organisms is their ability to reproduce and regenerate individuals with similar observable characteristics. This observation has led to scientists to wonder which component might be the cause of this inheritably maintained traits and how such important information is stored, interpreted, and transmitted by different individuals.

Thousands of studies have been conducted in the last century to try to discover the nature of this information and how it affects inheritance and evolution.

The first known observations about these hereditary traits date from 1866, when Gregory Mendel hypothesised about the existence of cell elements ("*Zellelemente*" in original German) as the responsible of the visibly inherited characteristics of the garden pea (*Pisum sativum*)(Burndy Library, Mendel and Punnett 1866). Since that moment, different authors, such as the botanists Carl Correns (Correns 1900), Hugo de Vries (Vries, Hugo de 1900) and Erich von Tschermark (Tschermak., E 1900) were able to reproduce Mendel's results, reinforcing the hypothesis of the existence of inheritable elements within the different plant species studied.

However, it was not until 1909 that Wilhelm Johannsen introduced both the concept of "gene", as a "unit of heredity", and the difference between the set of these units (genotype) and visible traits under study (phenotype) (Johannsen 1909).

This abstract and vague definition of gene has evolved to a more widely accepted version, in which a gene is a deoxyribonucleic acid (DNA) sequence that specifies for one or more sequence-related ribonucleic acids (RNAs) and/or proteins (Portin and Wilkins 2017). The complete set of genes in an organism is the genome. Hence, genetics, will be, broadly, the study of all aspects of genes, whereas genomics will be the study of the whole genome of organisms.

Although it was Mendel who first set the foundations of modern genetics, the research conducted by hundreds of scientists in the last century have enriched the knowledge in the field, and enable the development of techniques that allowed important discoveries in science, from the isolation and determination of DNA as the molecule responsible of the hereditary information to how it is related to observable traits in organisms.

DNA was first isolated by the Swiss scientist Friedrich Mieschner in 1869, when trying to isolate the protein components of leukocytes. During his experiments, he

isolated a substance with unexpected properties: it could be precipitated by acidifying the solution, and then re-dissolved in alkaline solutions, but not in water, acetic acid, diluted hydrochloric acid or solutions of sodium chloride, hence, not belonging to any of the known proteins. However, it was not until 1871 when Mieschner was able to develop a leukocyte nuclei extraction protocol to study the properties of this compound, which exhibited resistance to proteolysis and much higher content in phosphorus than the proteins described up to that moment. These characteristics lead him to conclude the discovery of a new substance, which he named “nuclein” (Miescher F. 1871; Dahm 2008).

The discovery of this new compound inside the nuclei of cells increased the interest in the study of this organelle, very poorly understood at that moment. Subsequently, in 1881 Albrecht Krossel identified nuclein as a nucleic acid, and proposed its current name, deoxyribonucleic acid (DNA), and also isolated the five nitrogen bases responsible of building the blocks of DNA and RNA: adenine (A), cytosine (C), guanine (G), thymine (T) and uracil (U) (Kossel A. 1881).

However, it was not until 1919 when Phoebus Levene proposed the “polynucleotide model”, describing the nucleotide components of nucleic acids from yeast nuclei (Levene P.A., 1919). Levene proposed that nucleic acids were composed of a series of nucleotides, and that each nucleotide was composed of one of four nitrogen containing bases, a sugar molecule, and a phosphate group, and these four bases were always repeated in the same order (Pray, L 2008).

However, the importance of DNA in the cell was not clear until the publication of Avery, MacLeod and McCarty results on its role in bacterial transformation, in 1944 (Avery, Macleod and McCarty 1944), where they suggested that it was DNA, and not proteins (as proposed up to that moment), the molecules in charge of transfer genetic information between bacteria, and therefore the responsible of storing the hereditary material.

Impressed by the studies Avery and collaborators, and with the model proposed by Levene, Erwin Chargaff (Chargaff *et al.* 1951) decided to apply recently developed chromatographic methods to study if there were any differences in DNA among different species, reaching two major conclusions: (i) The nucleotide compositions varies among species, which indicated that the nucleotides were not repeated in the same order, as Levene first proposed. (ii) Almost all DNA (no matter the organism or tissue under study) presents certain properties that are maintained, highlighting that the amount of adenine is usually similar to the amount of thymine, and the amount of guanine is similar to the amount of cytosine, which means that the total amount of purines (A+G) and pyrimidines

(C+T) should be equal. This phenomenon is currently known as the “Chargaff’s rule” (Pray, L 2008).

These results proposed by Chargaff, combined with DNA X-ray crystallography developed by Rosalind Franklin and Maurice Wilkins, contributed to one of the most recognisable breakthroughs in history of science: the proposal of the DNA structure by Watson and Crick in 1953 (Watson and Crick 1953). In their publication they emphasised in four fundamental features of the molecule: (i) The DNA molecule is a double helix, with two strands, connected by hydrogen bonds. In this structure, the complementarity between thymine and adenine and guanine and cytosine, is fundamental, as suggested in previous studies by Avery, MacLeod and McCarty (Avery, Macleod and McCarty 1944), and Chargaff (Chargaff *et al.* 1951) (ii) Most DNA double helices are right-handed, (only Z –DNA is left handed) (iii) The double helix is anti-parallel, which means that the 5’end of one strand is paired with the 3’end of the complementary strand, and the nucleotides are united by bonds between the 3’end phosphate of one sugar to the 5’end phosphate of the next sugar (iv) The outer edges of the nitrogen-containing bases are exposed and available for potential hydrogen bonds, which provides easy access to the DNA for other molecules (Pray, L 2008).

Although some minor changes have been implemented, these four major features of the model proposed by Watson and Crick are currently the same.

Therefore, a molecule of DNA is made of two molecular strands of nucleotides around each other, forming a double helix. There are four different types of nucleotides, and each one of them is formed by a deoxyribose sugar, a phosphate group, and a nitrogenous base. Although the sugar and the phosphate group are maintained constant in each nucleotide, the nitrogenous base is variable, and this variability conditions the nucleotide. The four different bases are adenine (A), thymine (T), guanine (G) and cytosine (C).

In each strand the sugar and the phosphate group are placed in the outside of the structure, whereas the internal part of the helix is occupied by the nitrogen bases, which bind by complementarity to the nucleotide present in the other strand by hydrogen bonds (adenine always binds thymine with two hydrogen bonds, whereas cytosine always binds guanine with three hydrogen bonds) (Griffiths 2012) (**Figure 1.1**).

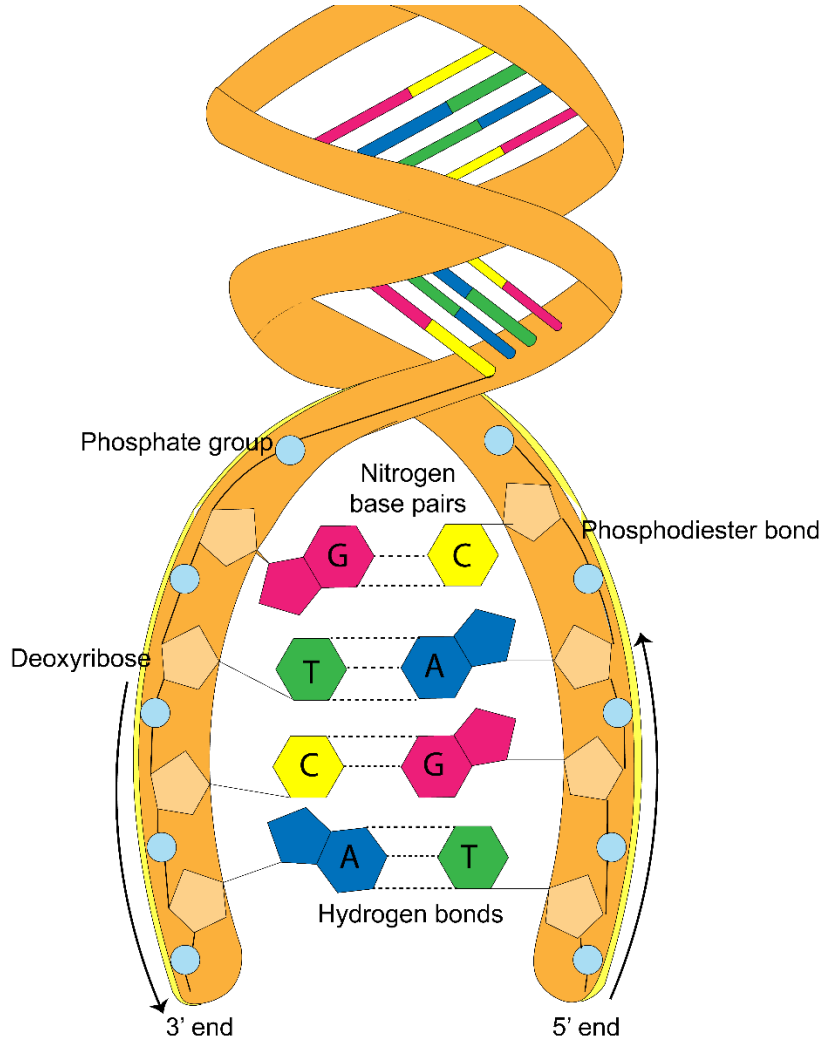


Figure 1.1 Components of the DNA molecule forming the double helix structure. Abbreviations: A: Adenine, C: Cytosine, G: Guanine, T: Thymine.

However, the role of the DNA molecule for the storage of genetic information relies not only on a code that is responsible for specifying the sequences of RNA and protein molecules and the sequence specific recognition sites for DNA-binding proteins, but also on a structural code that does not depend on the base sequences, but on their interactions.

The direct or indirect recognition of the DNA by the proteins is a major determinant of binding selectivity. In direct readout, the proteins surface interacts specifically with the DNA binding sequence, whereas in the indirect recognitions, the binding affinity will depend on the recognition of a specific DNA structure, such as a DNA bend or bubble, but does not always require a direct contact between protein and DNA (Travers and Muskhelishvili 2015).

The determination of the DNA structure by Watson and Crick was fundamental to demonstrate one of the most important requirements of any genetic material: its ability

for self-replication. However, the relation between DNA and proteins, and specially how the information was transferred from the nucleus to the cytoplasm, where proteins are sequenced, was unclear. In this context, Francis Crick proposed in 1958 the Central Dogma of Molecular Biology, in which he tried to summarise and explain the flow of genetic information within the cell. Based on the evidences, direct or indirect, he proposed 4 different transfer pathways: (i) DNA to DNA, (ii) DNA to RNA, (iii) RNA to protein and (iv) RNA to RNA (Crick 1958).

This theory, however, has been constantly challenged and updated as new discoveries provided information on how differently information is transferred, especially after the discovery of RNA reverse transcriptase, protein prions, protein folding by chaperones, or epigenetic modifications (Morange 2009). Currently, it is widely accepted that although the genetic information flow in the cell can go from DNA to DNA and RNA, and from RNA to RNA, DNA and proteins, translation to proteins is an irreversible phenomenon (Koonin 2015) (**Figure 1.2**).

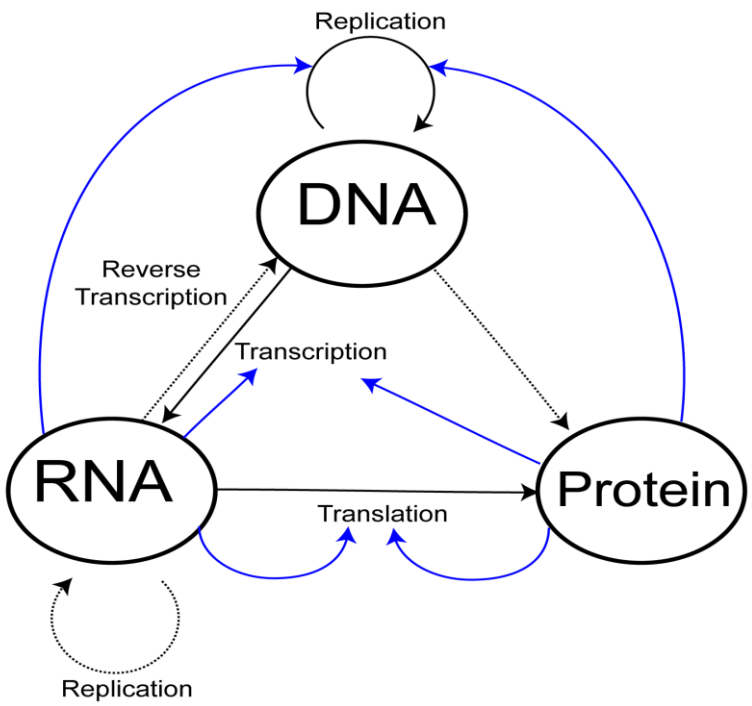


Figure 1.2. Current Central Dogma of Molecular Biology. Arrows indicate the flow of information. Black lines indicate processes known when the dogma was first proposed. Dashed lines indicate subsequent modifications. Blue lines indicate role of the molecules in regulation of the process.

One of the most important features any molecule susceptible to act as genetic material is the ability to maintain its number of copies through cell division and cell damage. The process in which DNA is able to copy itself is named replication, and for it,

the two strands of DNA separate from each other, and newly synthesised nucleotides are deposited onto the old strand, pairing each nucleotide with its complementary partner by a group of enzymes called DNA polymerases (Stillman 2008). To synthesise proteins, the cells must be able copy the molecule of DNA into another molecule called ribonucleic acid (RNA). This process is called transcription, and it is conducted by the RNA polymerases (**Figure 1.2**). The RNA produced is also composed of nucleotides, but the sugar is not deoxyribose, but ribose, and the thymine is replaced by uracil (Alberts *et al.* 2002). In eukaryotes this process is conducted in the nucleus of the cell, and once finished there are subsequent modifications in the molecule, such as the removal of introns (splicing), which will produce the final form of the RNA, the mRNA (messenger RNA). This molecule will be sent from the nucleus to the cytoplasm, where ribosomes will synthesise the proteins according to the information of nucleotides read in the mRNA, in a process called translation (Griffiths 2012).

The extensive progress observed in genetics and molecular biology towards the position it currently holds is importantly related to the use of model organisms in research. Model organisms have been a key factor for revealing cellular mechanisms and provide information about biological problems, mainly because they are useful for standardisation of results and are thought to be representative of larger class of individuals. Several organisms have been used as model organisms, such as *Escherichia coli* (bacteria), *Saccharomyces cerevisiae* (yeast), *Caenorhabditis elegans* (nematode), *Arabidopsis thaliana* (plant) or *Mus musculus* (mammal).

The importance of yeasts as model organisms was highlighted in 1996, when the complete genome of the budding yeast *Saccharomyces cerevisiae* (strain S288c) was published, being the first eukaryote completely sequenced (Goffeau *et al.* 1996; Mewes *et al.* 1997). This huge scientific milestone was reached thanks to the collaboration of numerous laboratories in Europe, North America, and Japan. After the publication, several libraries of strains have been developed in order to understand the genetics and cellular processes of eukaryotes, such as deletion libraries (Giaever *et al.* 2002), overexpression mutants (Sopko *et al.* 2006), or libraries with genes tagged with reporters (Huh *et al.* 2003).

Despite of this, the use of *S. cerevisiae* had been very extensive, especially in brewing industry, since the discovery of its role in alcoholic fermentation by Louis Pasteur in 1860 (Louis Pasteur 1860). The strain S288c was obtained by crossing several parental strains (Mortimer and Johnston 1986), and rapidly became the reference isolate, since it was able to maintain a very stable haploid genome, making the study the effects of gene mutations much simpler (Liti 2015).

Moreover, there are other features that make yeast an important model organism for research: (i) The genome is completely sequenced, which provides information for gene ontology studies and prediction of homologue genes function (Botstein, Chervitz and Cherry 1997) (ii) They are unicellular, but present different cell types (such as haploid MATa, haploid MATα, diploid MATa/α or spores) and the cell cycle and processes that differentiates them are known (Haber 2012) (iii) Their rapid proliferation makes them very cost effective, since a large amount of biomass can be produced in short time (Hanson 2018) (iv) As an eukaryotic organism, present molecular processes that are representative of complex eukaryotes (such as posttranscriptional modifications, or specific transcripitional regulation systems) (Lodish H., Berk A and Zipursky SL 2000) (v) Techniques that allow genetic modifications and cell biology studies have been extensively studied, developed and optimised (Hanson 2018).

For these reasons, *S. cerevisiae* is one of the most widely used eukaryotic models, with applicability for studies in aging (Murakami and Kaeberlein 2009), regulation of the gene expression (Biddick and Young 2009), signal transduction (Hohmann, Krantz and Nordlander 2007), cell cycle (Nasheuer *et al.* 2002), metabolism (Quirós *et al.* 2013), apoptosis (Owsianowski, Walter and Fahrenkrog 2008), and other biological processes (Karathia *et al.* 2011).

1.2 GENOME EVOLUTION IN YEASTS

The publication of the complete genome of the budding yeast *Saccharomyces cerevisiae* in 1996 (Goffeau *et al.* 1996; Mewes *et al.* 1997) revolutionised several fields in biology, being evolutionary biology one of them. Since then, comparative genomics expanded the scope of molecular evolution.

Normally two main types of studies were conducted: studies based in the comparison of sequences that were so extremely closed that they could be aligned at nucleotide level, or studies of genomes that were more distant, but close enough to find clear similarities (Zarin and Moses 2014).

Nevertheless, the current availability of new genome sequences allows not only the measurement of the evolution of thousands of genes using statistical analysis, but also new type of analysis in functional genomics, such as non-coding DNA studies, expression patterns and regulatory networks, post-translational modification or protein interactions, between multiple species, heralding the new era of population genomics.

Population genomics can be broadly defined as the use and analysis of sequencing data from genome wide loci of a large number of individuals from the same species, in order to understand better evolution (Peter and Schacherer 2016). The importance of exploration of interspecific genetic diversity lies on the information obtained about: (i) demographics, relationships and evolutionary history of populations, (ii) evolutionary processes involved in the generation and maintenance of genetic diversity, which are the forces that shape genome architecture (iii) the importance of different adaptive molecular variations for fitness (Luikart *et al.* 2003) (iv) distinction between effects that act at whole genome level, such as drift or migration, and those which act on individual loci, such as mutation or recombination (v) relationships between genotypes and phenotypes, especially for tracking allelic variants that are responsible for phenotypic divergence (Peter and Schacherer 2016).

Evolution-related population studies have been conducted in organisms from all kingdoms of life, starting from model organisms, such as *Saccharomyces cerevisiae* (Schacherer *et al.* 2009)*, Arabidopsis thaliana* (Cao *et al.* 2011)*, Caenorhabditis elegans* (Andersen *et al.* 2012) or *Drosophila melanogaster* (Mackay *et al.* 2012). This has helped to highlight the importance of yeasts as a model organism, whose main advantages lie on the power and relative ease of experimental design where the evolutionary history of molecular function in genes can be reconstructed and the fitness of populations directly measured (Dean and Thornton 2007). Despite of the lack of strong clear phenotypes like the found in other animal or plant model organisms, yeasts have been used to study the evolution of complex phenotypes, such as multicellularity (Ratcliff *et al.* 2015), transcriptional memory (Sood and Brickner 2017), or the mating locus switching (Hanson, Byrne and Wolfe 2014).

Furthermore, the cost and ease of genome sequencing and assembly have made even more favourable their role as model organisms for population genetics (Hittinger 2013). Several studies have been conducted comparing genomes of individual strains that have been helpful to characterise features of yeasts at the genome-wide scale (Schacherer *et al.* 2009; Hittinger *et al.* 2010). For example, Peter et al. (Peter *et al.* 2018) provided an accurate evolutionary picture of the genomic variations that shape phenotypes by the sequencing and phenotyping of 1.011 *S. cerevisiae* strains, proving a main difference in the nature of genomic variations between wild (SNPs) and domesticated (ploidy, aneuploidy and genome content) strains.

In the context of studying relationships between genotype and phenotype variations, population genomics have offered great improvements in tracing quantitative trait loci (QTLs) mapping in yeasts. QTLs allow the link of certain phenotypes with regions of the genome. Currently, population genomics allows the study of traditional

traits of interest (such as high temperature fermentation (Wang *et al.* 2019), or cell shape (Nogami, Ohya and Yvert 2007)), but also traits related with genetic variation, which help understand the complexities of inheritance and evolution in eukaryotes (Anderson *et al.* 2010).

The number of sequenced genomes is continuously increasing and that has allowed the development of genome-wide association studies (GWAS) in multiple organisms. However, there is a relatively low number of sequenced genomes from natural isolates of *S. cerevisiae*, which complicates the development of GWAS. Also, the majority of these genomes are from haploid or diploid strains that have been derived from single spores, biasing the natural ploidy and heterozygosity of isolates (Peter and Schacherer 2016). Despite of this, several attempts have been conducted to establish *S. cerevisiae* GWAS, such as the previously cited study Peter *et al*. (Peter *et al.* 2018), or the relationship between the genetic background of 165 *S. cerevisiae* strains and their mechanisms for toxins tolerance proposed by Sardi *et al*. (Sardi *et al.* 2018).

1.2.1 Duplications as a force driving evolution

The effects of gene duplications in evolution was first suggested in 1936, with Bridges (Bridges 1936) observations on the studies conducted by Sturtevant and Morgan (Sturtevant and Morgan 1923) and Sturtevant (Sturtevant 1925) about a gene duplication, detected by the doubling of a band in a chromosome, in a mutant of *Drosophila melanogaster* that exhibited extreme reduction in eye size. From there, several scenarios on how evolution was affected by gene duplication were proposed (Stephens 1951; Ohno 1967; Nei 1969) but it was not until 1970 with the publication of the book “Evolution by Gene duplication” by Susumu Ohno (Ohno 1970) that the topic was popularised among biologists.

The most obvious contribution of gene duplication to genome evolution is providing new genetic material that is susceptible of selection. This gives organisms enough genome plasticity to adapt to environmental changes and survive in the new conditions with possible new functions.

However, the prevalence and importance of gene duplication on evolution was not demonstrated until the late 1990s, when the genome sequencing techniques allowed the determination and analysis of complete genomes.

One of the most important contributions to evolutionary biology was published shortly after the release of the complete genome sequence of *S. cerevisiae* in 1996. Kenneth Wolfe *et al*., that had been involved in the *S. cerevisie* sequencing project, found 55 large duplicated gene blocks, which they considered as an indication of an ancient

complete genome duplication (Wolfe and Shields 1997). These duplicated blocks had particularities that seemed to indicate they were quite antique. First, the amino acid sequence between the gene pairs was only 63% similar, indicating high mutation rate, and second, despite of the block duplication, only around 25% of the genes present were actually duplicated, which seems to indicate that the rest had been duplicated and lost. The fact that there were almost no overlaps between the blocks and that their orientation respect to the centromeres and telomeres was conserved seemed to indicate that these were the result of a duplication event, that had after been rearranged by reciprocal translocations between the chromosomes (Wolfe 2015).

The hypothesis of this whole genome duplication (WGD) in *S. cerevisiae* was confirmed in 2004 with the publication of the genome sequence of several species that had branched off from this lineage before the WGD, such as *Ashbya gossypii* (Dietrich *et al.* 2004)*, Kluyveromyces waltii* (Kellis, Birren and Lander 2004), or *Debaryomyces hansenii* (Dujon *et al.* 2004) among others. These genome (or sometimes chromosomes) duplications occur probably by a lack of disjunction among mother-daughter chromosomes after DNA replication during cell cycle (Zhang 2003)

Although several examples of WGD events have been described in species across the eukarytic tree of life, from animals (Jaillon *et al.* 2004) to plants (Jiao *et al.* 2011), less extensive duplications can also be detected and affect and drive evolution. Apart from genome duplications, there are different phenomenon that can cause gene duplication, such as unequal crossing over, or retropositions.

Unequal crossing overs normally generate tandem gene duplications, which are duplicated genes linked in a chromosome, and depending on the position of the crossing over, the duplicated region can contain a complete gene, part of it or even several genes. When there are complete or several genes, all the regions (introns, intergenic spacers) will also be in the duplicated genes (Zhang 2003). In *S. cerevisiae,* this mechanism is, for example, responsible of the maintenance of homogeneous repeats in the rDNA gene (Szostak and Wu 1980), or in the gene *CUP1*, which codifies for a copper and cadmium binding protein, responsible of the detoxification of metal ions and elimination of superoxide radicals (Welch, Maloney and Fogel 1990).

Retroposition occurs when a gene is transcribed to mRNA, subsequently retrotranscribed to complementary DNA (cDNA) and re-inserted in the genome, so in this case, introns, regulatory sequences, poly A tracts and flanking short direct repeats will not be part of the gene duplication. Since these regulatory sequences are not duplicated, the new duplicated gene lacks the necessary elements required for transcription and thus immediately becomes a pseudogene, although they can be expressed if the

insertion is downstream a promoter region in the genome. This re-insertion is normally random, thereby gene duplications created by retroposition are rarely linked to the original gene (Zhang 2003). Schacherer *et al*. (Schacherer *et al.* 2004) described this phenomenon as being responsible of the duplication of ATCase, previously observed by Roelants *et al*. (Roelants *et al.* 1995) as a consequence of a chromosome rearrangement in *S. cerevisiae*.

1.2.2 Evolutionary fate of duplicated genes

Upon a gene duplication, two major events can occur, it can be fixed, or it can be lost in the population. When the allele comprising a duplicate gene is neutral, it only has a small probability of being fixed in a diploid population (1/2N, being N the effective population size), thereby, many gene duplications will be lost. For those which are not lost instead, it will take 4N generations to consider them fixed in the population (Kimura 1991). Once fixed, the function of the gene will determine the long-term evolutionary fate of the duplication.

Gene duplication will generate functional redundancy. This is not always advantageous, since the mutations that might disrupt one of the genes are not deleterious and will not be removed from the population by selection. Hereby, the gene accumulating mutations will eventually lose its function or its expression will be stopped, so it will become a pseudogene, which in the long-term will be removed from the genome or will accumulate mutations and will not be longer identifiable. This pseudogenization process will occur in the first few million years after duplication when there is no selection pressure under the duplicated gene (Lynch and Conery 2000). Several pseudogenes, mainly derived from transporter functional genes, have been identified clustered in the subtelomeric regions of yeasts spanning the phylogenetic range of hemiascomycetes, such as *S. cerevisiae, Candida glabrata* or *Zygosaccharomyces rouxii* (Lafontaine and Dujon 2010).

The functional redundancy caused by gene duplication can on the other hand be beneficial since there are extra copies to be transcribed and more RNA and proteins products might be available. This is important when the product of a gene is very important or required in large amounts, for example in the case of the rDNA gene.

When this is the case, both duplicated genes might maintain the same function. This can be achieved by gene conversion or by strong purifying selection.

Under gene conversion, two paralogous genes will have very similar sequences and functions, which is referred as concerted evolution. This is normally achieved by non-reciprocal exchange of genetic material between homologous sequences, which

can be beneficial when mutations that diverge the gene function are removed, or harmful when the mutation is exchanged by the functional gene (Carson and Scherer 2009). Alternatively, by strong purifying selection, the mutations that modify the gene function are removed, so the genes are prevented from functional diversion (Zhang 2003).

Although these two approaches might seem similar, the difference can be easily distinguished by the examination of synonymous nucleotides differences in duplicated genes. Synonymous (or silent) mutations on the nucleotide sequence cannot be removed by selection, but they can by gene conversion, since this homogenises the DNA sequences regardless of changes in the amino-acid translation, whereas in the case of gene functionality maintained by selection, several non-synonymous mutations that do not affect the main function of the product might be observed (Zhang 2003). Studies on genome evolution have suggested that purifying selection is much more important in maintaining common functions of duplicated genes (Nei, Rogozin and Piontkivska 2000; Piontkivska, Rooney and Nei 2002), whereas gene conversion is only favoured in very restrictive occasions (Hurst and Smith 1998).

In other cases, the presence of extra copies of a gene are not particularly advantageous, in which case these copies are unlikely to be maintained stably in the genome (Nowak *et al.* 1997). However, both copies can be theoretically maintained when they offer advantages in some of their functions (subfunctionalization), which sometimes occurs quite quickly after the gene duplication (Zhang 2003). Some of these changes can be variations in the levels of gene expression or its localization, or even at protein level. A perfect example of this subfunctionalization is the case of the paralogous genes *BAT1* and *BAT2* in *S. cerevisiae*. They codify for the branch chain aminotransferases (BCATs) Bat1 and Bat2, which catalyse the synthesis of branched chain amino acids (valine, isoleucine, and leucine, or VIL). *BAT1* is highly expressed under biosynthetic conditions (primary nitrogen sources) in the mitochondria, whereas *BAT2* is expressed under catabolic conditions (VIL as a nitrogen source) in the cytosol (Colón *et al.* 2011; González *et al.* 2017).

An important outcome for evolution after a gene duplication event is the origin of a new function from it (neofunctionalization). In this case several changes are normally required in the gene or protein codified and it is important to try to understand the possible changes in the protein function and structure during evolution. *YAP7* is a gene that has been identified in several yeasts, such as *S. cerevisiae* or *C. glabrata*. It codifies for the transcription factor Yap7, which regulates the expression of *YHB1*, a nitric oxide oxidase. Yap7 is part of the Yap family, originated after the yeast WGD, which indicates that there are other paralog genes in the genome (ohnologues), such as *YAP5*, which is expressed as a response to iron stress (Merhej *et al.* 2015). These differences in the

function of genes after an event of duplication constitute a great example of neofunctionalization.

All these different scenarios that might occur after a gene duplication are subjected to different evolutionary forces that will determine functional divergence of the duplicated genes. For example, cases of subfunctionalization of gene expression in different locations might be related to neutral evolution without the action of positive selection (Zhang 2003). However, in the case of functional specialization and neofunctionalization two models have been proposed.

The first model is known as the Dykhuizen-Hartl effect (Dykhuizen and Hartl 1983), and it proposes that gene specialization does not require positive selection. In this case, after a gene is duplicated, random mutations will promote its fixation in the genome by selection, but these mutations will induce a change in the functions of the gene when there are environmental changes. Although it was firstly described in bacteria (Dykhuizen and Hartl 1980), examples of this adaptation have also been observed in eukaryotes, as is the case of the *PMR1* gene in *S. cerevisiae*, a gene in which several mutations give faster adaptability to high salt stresses (Bui *et al.* 2015).

The second model requires positive selection and it involves two other scenarios. In the first one, after a gene is duplicated, a few neutral mutations will create a new weakly active function in one of the genes, and positive selection will allow the accumulation of mutations that enhance this new function. This seems to be the case of *YAP7* in *S. cerevisiae* (Merhej *et al.* 2015). In the second scenario, the ancestral gene presents two functions. After the gene duplications each one for them will specialise and adopt one of the functions of the ancestral gene, which will be positively selected and improved during evolution. For example, the gene pair *SIR3* and *ORC1* have different functions in *S. cerevisiae*, but both activities are carried out by the same protein in *Saccharomyces kluyveri*, which lacks the WGD (Conant and Wolfe 2008).

Once the genes have reached enough level of specialization or neofunctionalization, they are likely to be maintained in the genome.

A summary of the events that might take part in the fixation of duplicated genes is present in **Figure 1.3**.

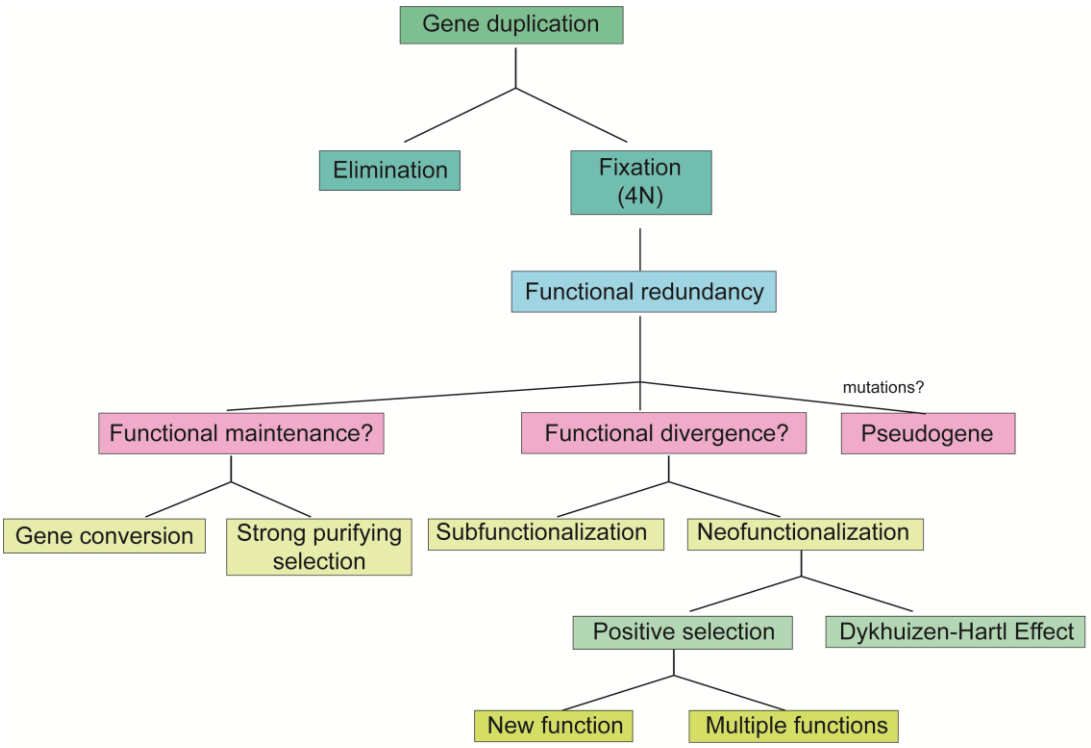


Figure 1.3. Fate of duplicated genes.

The study of the changes that the duplicated genes have acquired during evolution to obtain interesting new functions or improved expression profiles is remarkably important since this can give information about both which key amino acid changes are fundamental for functional divergence, and which changes could be interesting for a more targeted evolution in microorganisms used in different biotechnological processes.

1.2.3 Repetitive DNA in the genome

As previously discussed, one of the most obvious contributions of gene duplications to genome evolution is providing new genetic material susceptible of selection. However, genomic repeats can also affect evolution in different manners, as they might be hotspots for genome modifications associated to instability.

There are two main reasons behind this. First, repetitive regions are prone to recombination, and therefore they might produce insertions, deletions or translocations between adjacent or distant repeats (Argueso *et al.* 2008; Hoang *et al.* 2010), and even between imperfect repetitive sequences (Mézard, Pompon and Nicolas 1992). Second, non-canonical DNA structures are commonly observed in repetitive loci, which can lead to a collapse of the replication fork and to double strand breaks (Aguilera and García-Muse 2012).

The effects of repetitive regions on genome modifications related to instability have been studied in several yeasts, such as *S. cerevisiae* (Foss *et al.* 2017), and specially in yeasts belonging to the CTG clade, such as *C. albicans* (Freire-Benéitez, Gourlay, *et al.* 2016; Freire-Benéitez, Price, *et al.* 2016; Robert T Todd *et al.* 2019), a common human pathogen in which genome modifications associated to instability events have been related to its rapid adaptation to physiological changes in the host (Larriba G 2004).

Hence, the study of repetitive regions can also be important as a marker for genome evolution. Some of the most important contributors to genome repetitions in yeasts are transposable elements, telomeres, centromeres and mini- and micro-satellites.

1.2.3.1 Transposable elements

Transposable genetic elements (TEs) include a diverse group of DNA sequences with the ability to move to new sites in the genome either by themselves or by hijacking active TE machinery (Fedoroff 2012). The mobilization of TEs in the genome is called transposition or retrotransposition, depending on the nature of the intermediate required for the process (Munoz-Lopez and Garcia-Perez 2010).

Although they were first described as a genetic oddity when discovered in maize plants by Barbara McClintock (McClintock 1950), the presence of TEs has also been described in eukaryotic genomes (Cameron, Loh and Davis 1979; Adams 2000; International Human Genome Sequencing Consortium 2001; International Rice Genome Sequencing Project 2005).

TEs were first classified by Finnegan (Finnegan 1989) according to their transposition intermediate: RNA (class I, or retrotransposons, in which RNA acts as the transposition intermediate) or DNA (class II, or DNA transposons, where DNA is the transposition intermediate). However, the discovery of highly reduced non-autonomous TEs (also called miniature inverted repeat transponsable elements, or MITEs) (Bureau and Wessler 1992) and of TEs with variations in the transposition intermediate and mode of replication (Lai *et al.* 2005; Morgante *et al.* 2005), challenged the two-class system, obliging to a new classification in favour of enzymological categories. Wicker *et al*. (Wicker *et al.* 2007) proposed a new TE classification system based on the used for phylogenetic studies: class, subclass, order, superfamily, family and subfamily and insertion (in hierarchical order) (**Figure 1.4**).

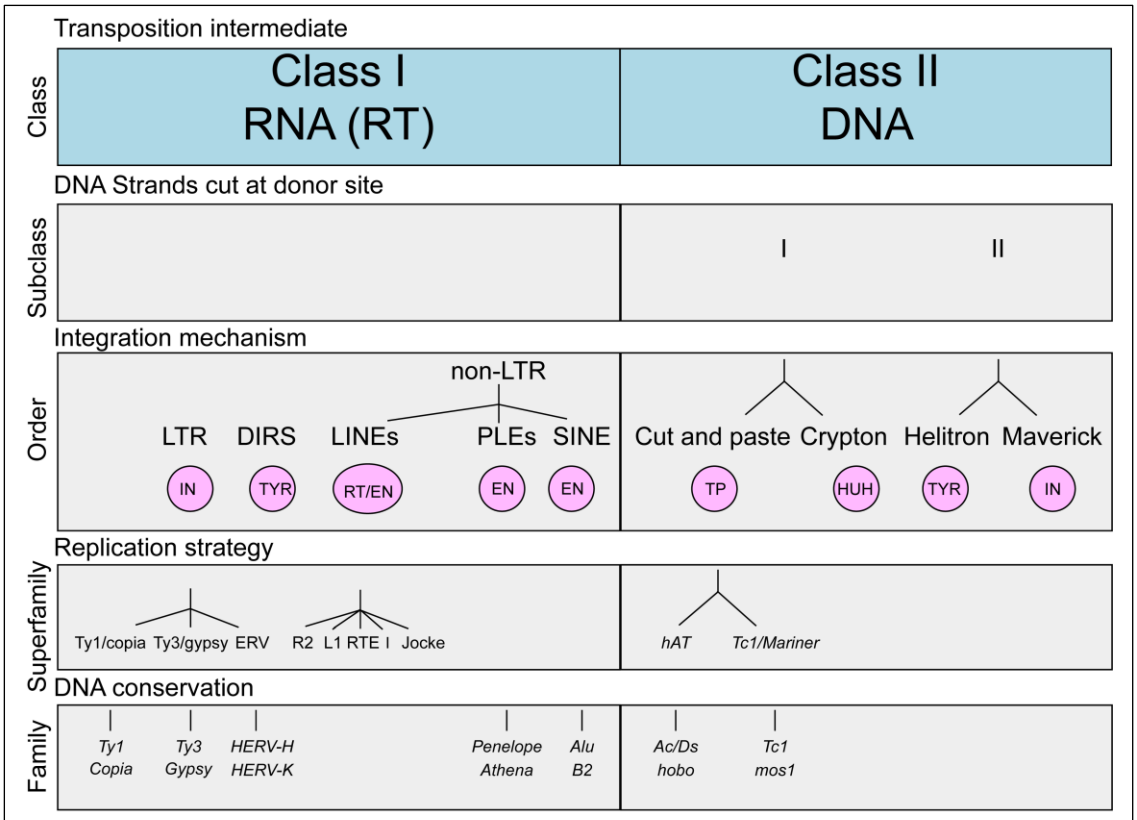


Figure 1.4. Eukaryotic transposable elements classification according to Wicker *et al*. (2007) with modifications by Bourque *et al.* (2018) and Makalowski *et al.* (2019). Purple circles indicate the enzymes involved in transposition: IN: Integrase, TYR: Tyrosine recombinase, RT: Reverse transcriptase, EN: Endonuclease, TP: Transposase, HUH: Rep/Helicase protein with HUH endonuclease activity, IN: Integrase. LTR: Long terminal repeats, LINEs: Long interspersed elements, PLEs: Penelope-like elements, SINEs: Short interspersed elements.

Therefore, this current classification divides TEs in class according to the presence of an RNA (class I) or DNA (class II) transposition intermediate (as the previous division). Then, subclass is used to differentiate between elements that copy themselves for insertion from those that leave the donor site for insertion elsewhere (so it also indicates the number of DNA strands that are cut at the donor site). Order is used to differentiate between the TEs according to their insertion mechanism, and in consequence, to their enzymology. Superfamilies share a replication strategy, but differ in other large-scale features, such as the structure of the protein, the non-coding domains, or the presence and size of the target site duplication (TSD). Therefore, there is no sequence conservation at DNA or protein level.

Superfamilies are then divided into families according to the DNA sequence conservation. Hence, similarity at the protein level is generally high between different families that belong to the same superfamily, but DNA sequence conservation is minimal, and normally restricted to highly conserved parts of coding regions. Subfamilies are defined on the basis of phylogenetic data, and might serve to distinguish between internally homogeneous autonomous and non-autonomous populations. Finally,

insertions describe one individual copy, corresponding to a specific transposition and insertion event.

1.2.3.1.1. Class I TEs

Class I TEs are characterised because they transpose via an RNA intermediate. In this case the DNA is transcribed to RNA and this is reverse transcribed to DNA by a TE-encoded reverse transcriptase (RT). Hence, the DNA is not cleaved at the donor site, and therefore there are no subclasses. Since each replication cycle will subsequently produce a new copy of the gene, these retrotransposons are also known as copy-and-paste, and are considered a major contributor to the repetitive fraction in large genomes (Bourque *et al.* 2018).

Retrotransposons can be divided into five orders according to their organization, mechanistic features and reverse transcriptase phylogeny: Long terminal repeat (LTR) retrotransposons, DIRS-like elements, *Penelope*-like elements (PLEs), Long interspersed elements (LINEs) and short interspersed elements (SINEs) (Wicker *et al.* 2007).

Upon integration, LTR retrotransposons produce a TSD of 4-6 bp. They typically contain open reading frames (ORFs) for gag, a structural protein for virus-like particles, and for pol, although an additional ORF of unknown function can also be found. POL encodes for an aspartic protease (AP), reverse transcriptase, RNase H (RH) and DDE integrase (INT). They also contain specific signals for packaging, dimerization, reverse transcription and integration (Wicker *et al.* 2007). Their size might range from a few hundred base pairs to more than 5KB. They are also characterised by a start sequence of 5‘-TG-3’ and an end sequence of 5’-CA-3’.

DIRS-like retrotransposons contain a tyrosine recombinase instead of an INT, so they do not form TSDs, and their terminal regions are unusual, since they resemble split direct repeats (SDR) or inverted repeats. All these suggests that their mechanism of integration is different (Wicker *et al.* 2007). Although they are present in higher eukaryotes and in fungi, they are absent from the genome of several model organisms, such as mammals, *S. cerevisiae* or *D. melanogaster* (Piednoël *et al.* 2011).

PLEs encode a RT that is more closely related to telomerase than to the RT from LTR retrotransposons or LINEs, and an endonuclease that is related to both intron-encoded endonuclease and to the bacterial DNA repair protein UvrC.

LINEs lack LTR, can reach several kilobases in length and are found in all eukaryotic kingdoms. They are divided in five superfamilies: *R2, L1, RTE, I and Jockey*, and each one of these superfamilies is divided in several families.

Autonomous LINES codify for at least a RT and a nuclease in the pol ORF, and only members of the subfamily I present RNaseH activy. They form TSD upon insertion, but this insertion is often truncated, which makes them difficult to find. The truncation is probably related to premature termination of reverse transcription. However, they are characterised by the possible presence of either a poly(A) tail or an A-rich region in their 3' end.

Finally, the SINE order is characterised by the presence of a Pol III promoter in the 5' end, which allows them to be expressed, and a variable 3' end, with either tandem repeats or a poly(T) tail (the Pol III termination signal). The presence of this promoter is explained by the origin of the TE: the accidental retrotransposition of different polymerase III (Pol III) transcripts. Although they can express, they are non-autonomous, since they rely on LINEs for transposition functions, such as RT. They are normally small (80-500 bp) and generate TSDs (5-15 bp).

1.2.3.1.2. Class II TEs

Class II TEs are characterised because the transposition intermediate is DNA, and they can be divided in two subclasses, depending on the number of DNA strands cut during transposition. Subclass I are the TEs that suffer a double stranded DNA cleavage, whereas subclass II undergo transposition without double stranded cleavage (Wicker *et al.* 2007).

The subclass I-Class II TEs are characterised by the presence of terminal inverted repeats (TIRs) of variable length, which are recognised by a transposase. The transposase will cut both strands of DNA at each end. These TIRs and the TSD size are used to differentiate between the 9 superfamilies in which this subclass is divided. The most common and better described superfamily is the *Tc1-Mariner*, which is ubiquitous in eukaryotes and is characterised by a simple structure of two TIRs and a transposase ORF (Wicker *et al.* 2007).

The number of subclass I-Class II TEs can increase during evolution, since they can transpose to different part of the genome during DNA replication, for example, moving from a position that has been already copied to another one where the replication fork has not passed yet, or by exploiting the gap repair machinery after excision from the donor site (Wicker *et al.* 2007).

The subclass II-Class II TEs undergo transposition without double stranded cleavage, and therefore they are known as cut-and-paste TEs. This subclass is divided in different orders, such as *Helitron* or *Maverick* (Wicker *et al.* 2007). In the case of *Helitron* the transposition seems to be regulated by a replication via a rolling-circle mechanism, with only one strand cut, generating a circular DNA intermediate that will not generate TSDs after insertion (Grabundzija *et al.* 2016).

1.2.3.1.3. Contribution of TEs to genome evolution

The ability of TEs of changing their position over time is one of the main reasons of their importance in genome evolution (**Figure 1.5**). However, this transposition is far from being random, and there are several factors that condition TEs behaviour.

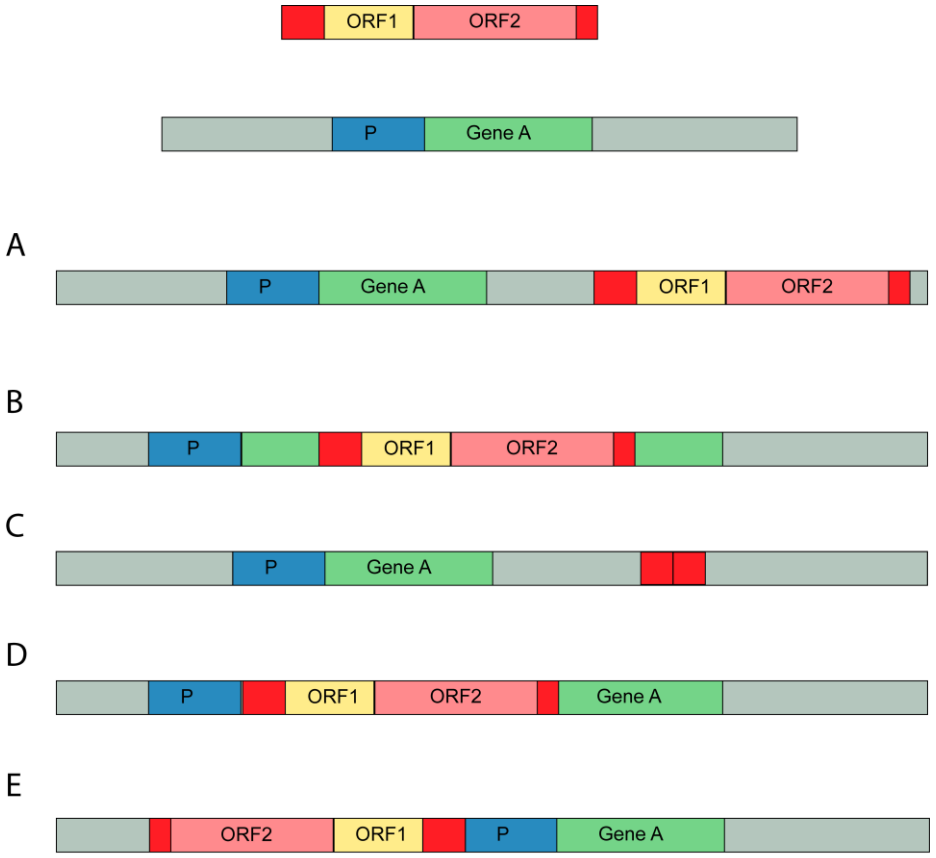


Figure 1.5. Effects of transposable elements (TEs) upon integration. Integrations are normally in non-coding regions, or regions which will not cause deleterious effects in the cell (A). However, TEs can cause gene disruption by insertion in a coding region (B). When they transpose the excision is not always perfect, leaving repetitive elements behind, which might cause genome instability effects (C). Finally, they might change gene expression by regulating the expression of the gene itself (D) or the promoter (E) also when inserted in the opposite orientation.

TEs exhibit various levels of preference for insertion. Normally they select regions that will facilitate their future propagation, but at the same time will not compromise the survival of the host (**Figure 1.5, A**). The genes that are present in different regions of the genome might be inhibited upon TEs insertion, so they act as an important source of mutagenesis (**Figure 1.5, B**). This might lead to non-beneficial mutations, and their disappearance, so several TEs have evolved to target genes with no important effects in the cell survival, such as the RNA polymerase III, where they retain the ability to be transcribed, but do not affect the host gene expression (Bourque *et al.* 2018).

Another important force in shaping the distribution of TEs along the genome is natural selection, since insertions strongly deleterious will be removed from the

population, whereas insertions with no, or little, effect on the host might be fixed, and even accumulate mutations during further evolution processes, some of which might even inactivate the transposition capacity, hence, becoming inactive (Lynch. M 2007).

Normally, they do not offer benefits through mutations, but they persist in evolution, not only through vertical transmission from parent-to-daughter cell, but also by horizontal transmission between individuals and even species (Wallau, Vieira and Loreto 2018).

However, to persist in evolution, TEs need to balance between expression and repression. Expression must be high enough to promote amplification, but it should not generate fitness disadvantages to the cells. There are some evolutionary strategies that TEs have developed to control their expression/repression balance, such as self-regulatory mechanisms (Lohe and Hartl 1996; Saha *et al.* 2015), enzymes with suboptimal activity for transposition (Lampe *et al.* 1999; Mátés *et al.* 2009), or insertion in transcriptionally inactive chromatin (heterochromatin) (Brady *et al.* 2008). Hence, other cellular mechanisms are important in the control of the balance, such as chromatin and DNA modification pathways (Goodier 2016).

Although the most obvious effect of TEs on genome evolution might be the disruption of genes in the region of insertion, they have different ways to damage the cell. For example, the action of different TE-encoded proteins can lead to genome instability (Hedges and Deininger 2007), they might interfere with transcription (Elbarbary, Lucas and Maquat 2016), or through the accumulation of RNA transcripts or extrachromosomal DNA, which might lead to defence responses in cells (Kassiotis and Stoye 2016), such as the RNAi system used as defence against nucleic acid molecules, and that has been described to target TEs in yeasts like *S. pombe* (Hansen *et al.* 2005) or *Cryptococcus neoformans* (Janbon *et al.* 2010).

Their role as cause of genome instability also helps to shape evolution through several changes in chromosome features. Their insertion and cleavage is not always precise, so the regions surrounding them can be affected by either disruption or by movement to other regions with transposition, which will generate repetitive sequences in the genome (Bourque *et al.* 2018) (**Figure 1.5, C**). These microhomology regions generated can cause template switching during the repair for replication errors, which is a source of structural variants (Bourque *et al.* 2018). Moreover, chromosome rearrangements can be induced by transpositions, especially after they have lost their capacity to mobilise (Carvalho and Lupski 2016). The main reason is that recombination events can occur between highly homologous sequences that have been dispersed by TEs at different positions in the genome (Bourque *et al.* 2018).

Moreover, there are other features in which TEs can condition the evolution of chromosomes. For example, in the genus *Drosophila*, LINE-like retrotransposons maintain the telomeres, since the telomerase has been lost during evolution (Pardue and DeBaryshe 2011), which also suggests, that the reverse transcriptase activity of telomerase has been originated from retroelements (Belfort, Curcio and Lue 2011). Furthermore, TE sequences and non-active transposons seem to play structural roles at centromeres in both budding (Coughlan and Wolfe 2019) and fission yeasts (Casola, Hucks and Feschotte 2007), but also in higher eukaryotes like mammals (Casola, Hucks and Feschotte 2007).

Their main contribution to regulate gene transcription upon insertion is related to the modification of *cis*-regulatory elements. Since TEs present their own transcription factors, they are able to modify expression networks of genes that are near to the insertion locus, or tune pre-existing networks (Bourque *et al.* 2018) (**Figure 1.5 D and E**).

Furthermore, it is important to note that all these contributions can be also increased when cells are under stressful conditions, since transposition seems to be upregulated in this case for a number of organisms (Horváth, Merenciano and González 2017; Lanciano and Mirouze 2018).

1.2.3.1.4. TEs in yeasts

Yeasts present a wide interspecific diversity of TEs, with substantial differences in TE content, families, and distribution. Class-I are often found, whereas Class-II have been described in a very limited number of species and have only been recently identified (Bleykasten-Grosshans and Neuvéglise 2011).

The yeast genome is characterised as having a low content of TEs (less than a 5%) (Wöstemeyer and Kreibich 2002). The first transposon identified in yeast was a Class-I Ty element that was bearing LTRs in *S. cerevisiae* (Cameron, Loh and Davis 1979), whereas both non-LTR (Chibana *et al.* 1998) and Class-II elements were first identified in *C. albicans* (Goodwin, Ormandy and Poulter 2001). Since then, different types of TEs have been identified in yeasts, with special prevalence of LINEs (Bleykasten-Grosshans and Neuvéglise 2011).

Some species are considered to be TEs 'reservoir', since they contain a high number of TEs, such as *S. cerevisiae,* or because their TEs are potentially active and are from diverse classes and families, which is the case of *C. albicans, Yarrowia lipolytica,* or *C. neoformans*. However, other species are 'empty' of TEs, whether because they are not active (*Eremothecium gossypii,* or *Zygosaccharomyces rouxii*), or

because there have been no traces of them so far (*Pichia sorbitophila*) (Bleykasten-Grosshans and Neuvéglise 2011).

There are several studies that try to connect this variability in the number of TEs with the evolutionary history of the organism, since this might accelerate their expansion or elimination. It has been noted that 'reservoir' species seem to be pathogenic for humans or plants (such as *C. albicans, C. dubliniensis,* or *Phytophtora infestans*), which suggest that their presence and contribution to mutagenesis and genome instability might be an advantage for adaptation. However, the presence of TEs is not a fundamental feature for pathogens, since certain species, such as *C. glabrata* lacks TEs (Bleykasten-Grosshans and Neuvéglise 2011).

Phylogenetically, the origin of yeast TEs is diverse. The presence of *copia*-like LTR transposons (or Ty1) in *Saccharomycetaceae* suggest that they all descended from a common ancestor that evolved in the different lineages by acquisition of the structural features that characterise them (gag and pol ORFs), which have then evolved independently (Bleykasten-Grosshans and Neuvéglise 2011). Others, such as the *hAT*-like, the *Tc1-Mariner* and the *Mutator* elements seem to have been acquired by horizontal transfer in *Lachancea* and *Yarrowia* species or in *C. albicans* (Neuvéglise *et al.* 2005; Génolevures Consortium *et al.* 2009).

Apart from the wide interspecific diversity offered by yeasts TEs, there are also studies that demonstrate the high variability in number and location within strains of the same species. This intraspecific diversity is so important that it is considered one of the richest sources of generic variability, altogether with SNPs and subtelomeric variations (Carreto *et al.* 2008). However, the identification and annotation of the exact position of TEs is difficult given their repetitive nature and the impossibility of the genome assemblers to map them correctly, which complicates the study of intraspecific TEs (Bleykasten-Grosshans and Neuvéglise 2011).

Despite of this, several methodologies have been developed to successfully map TEs, for example by combination of microarrays and sequencing (Gresham, Dunham and Botstein 2008), and even the study of their potential activity is possible by identification of their insertion sites (except of movements by homologous recombination between two homologous non-active TEs) (Bleykasten-Grosshans and Neuvéglise 2011).

Mapping the position of TEs and their possible integration sites has allowed the study of their impact on genomes, since different signatures can be associated with differently evolved strains. For example, *S. cerevisiae* strains used for wine fermentation present less Ty content (Dunn, Levine and Sherlock 2005) and an inversion in the

Ty1/Ty2 ratio (Novo *et al.* 2009) when compared to the reference laboratory strain S288C.

Chromosomal rearrangements are one of the most important impacts of TEs to the evolution of yeast genomes. The high number of repetitions provided by TEs will be involved in several phenomena, such as homologous recombination, which might be the cause for TE polymorphisms (presence of full length TEs, or the repetitions associated to them), deletions, duplications, inversions or translocations (Lemoine *et al.* 2005) among other spontaneous mutations (Lynch *et al.* 2008).

Several studies have reported that Ty-related chromosomal rearrangements (mainly duplications) increase the fitness of strains of *S. cerevisiae* growing under strong selective pressure (Maitreya J. Dunham *et al.* 2002; Libuda and Winston 2006; Gresham *et al.* 2010; Demeke *et al.* 2015), which suggests that they facilitate their evolution (Bleykasten-Grosshans and Neuvéglise 2011). Moreover, their action as a gene promoter in yeasts has also been reported, which means that they can regulate the expression of the genes adjacent to their insertion site (Chisholm and Cooper 1992; Servant, Pennetier and Lesage 2008).

In conclusion, TEs can shape the genome of yeasts by generating genomic variability, through all gene mutagenesis, chromosomal rearrangements or transcription modifications, which might generate beneficial phenotypes during adaptation to environmental stresses, or new niches.

1.2.3.2 Telomeric repeats

The DNA replication machinery cannot copy the very end of linear chromosomes. However, coding information is not lost owing to the presence of non-coding repetitive DNA sequences: the telomeres. Hence, the main function of telomeres is to prevent chromosome shortening and end recognition by the DNA repair machinery, and avoid destabilizing chromosome end-to-end fusions (Blackburn 1991; Stewart *et al.* 2012).

Telomeres are comprised of double-stranded and single-stranded DNA sequences, normally consisting in G-rich tandem repeats. These repeats are very diverse, and several organisms from protozoa to higher eukaryotes present irregular telomere consensus sequences, such as $TG_{1\text{-}4}G_{2\text{-}3}$ for *S. cerevisiae* (Shampay, Szostak and Blackburn 1984; Wang and Zakian 1990), $GGTTACA(G)_{1\text{-}4}$ in *S. pombe* (Hiraoka, Henderson and Blackburn 1998), $ACG_2ATGTCTA_2CT_2CT_2G_2TGT$ in *C. albicans* (McEachern and Hicks 1993) or T_2AG_3 for humans (Moyzis *et al.* 1988).

The length of telomeres varies across several species, from 300-400 repeats in yeasts to tens of thousands in higher eukaryotes (Vega, Mateyak and Zakian 2003).

However, their protective role over coding sequences results in the decrease of the number of protective telomeric repeats with each DNA replication and cell division cycle, since they cannot be restored by the DNA replication machinery (Watson 1972; Olovnikov 1973). There are two main strategies used to maintain the stability of the telomeres: telomerase dependent or telomerase independent mechanisms.

Telomerase is a ribonucleoprotein complex, which is recruited to short telomeres to add telomeric DNA repeats to the chromosome ends (Teixeira *et al.* 2004). The components of the complex can vary across species, but telomerase reverse transcriptase (TERT) and telomerase RNA component (TERC, or TLC1 in yeasts) are common to all (Hall *et al.* 2017). TERC will act as a template for reverse transcription and provides a conserved catalytic core required for scaffolding of all the telomerase protein components to form the telomerase RNP holoenzyme (Niederer and Zappulla 2015).

Conversely, small populations of telomerase-null yeast cells are able to maintain telomere elongation using Rad52-depedent pathways, type I and type II, collectively known as RTE (RAD52-dependent recombinational telomere elongation) (Hall *et al.* 2017). In the type I pathway the telomeres are extended by the acquisition of subtelomeric Y' sequences, which are elements that share features of degenerate transposable elements. The movement of these Y' elements is normally by recombination between them and a deprotected telomere, and depends on the proteins Rad52, Rad51, Rad54, Rad55 and Rad57 (McEachern and Haber 2006). In the type II pathway, there is an elongation of the terminal telomeric repeat tracts. This elongation depends on several proteins, apart from Rad52, such as Rad59 or the MRX complex. Some studies suggest that both Rad59 and the MRX complex promote the annealing of a single-stranded telomeric overhang and another telomeric terminus by recombination (McEachern and Haber 2006).

Telomeres length is ligated to cell death. Therefore, to avoid immortality of the cells, the activity of telomerase must be controlled. Telomerase recruitment, access to telomeres, and frequency and extent of the telomere elongation is regulated *in cis* by the activity of both single and double-stranded DNA-binding proteins (Cifuentes-Rojas and Shippen 2012), such as Rif1, Rif2, or Rap1 (Hass and Zappulla 2015; Luo, Vega-Palas and Grunstein 2002; Kaizer *et al.* 2015; Hall *et al.* 2017).

These capping factors present also other functions to ensure telomere stability. Since telomeres are long repetitions of DNA sequences, they could be detected by the DNA repair machinery as a DNA break and fused to each other. To prevent this fusion between telomeres on different chromosomes or chromosome arms to each other

several telomere capping factors (such as Rap1 in *S. cerevisiae*) ensure the local inhibition of non-homologous end-joining (NHEJ) DNA repair pathway (Marcand *et al.* 2008).

1.2.3.3 Centromeres

Centromeres are regions in the chromosome classically defined as constrictions that appear during the metaphase, and that are in charge of holding sister chromatids together and of binding to spindle microtubules to mediate in the separation of chromosomes to daughter cells during anaphase (Roy and Sanyal 2011; Yadav *et al.* 2018).

Although the presence of centromeres is important for correct cellular functioning, their structure is species specific, and there is no common sequence that describes the centromeres of all organisms, which is an indication on how fast these sequences evolve and how they can condition speciation (Malik and Henikoff 2009).

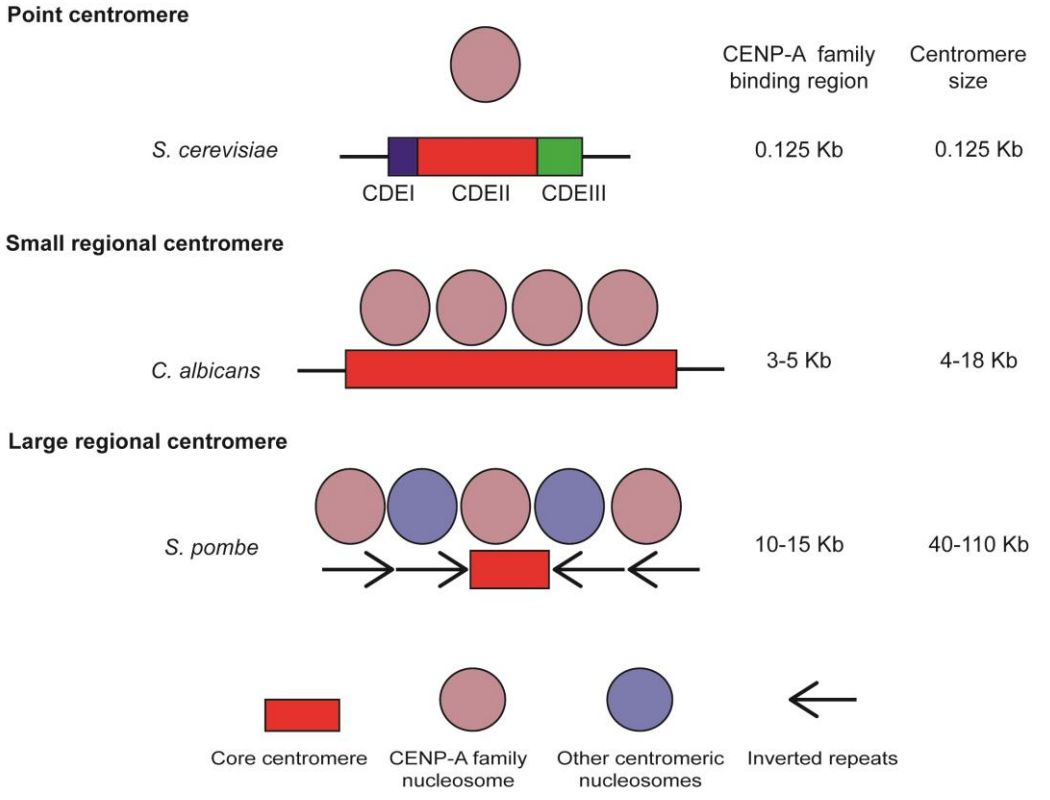


Figure 1.6. Organization of DNA elements and nucleosomes in point, small regional and large regional centromeres (Adapted from Roy and Sanyal, 2011). Abbreviations: CDE: centromeric DNA elements.

Despite this diversity, there is one common feature in all the functional centromeres: the presence of specialised chromatin marked by the centromere-specific histone H3 (CenH3) variant of the CENP-A family: Cse4 in *S. cerevisiae*, Cnp1 in *S. pombe*, CID in *D. melanogaster* or CENP-A in humans (Roy and Sanyal 2011). These proteins determine the centromere identity, as they seed the formation of functional

centromeres across organisms. Moreover, they condition the formation of the kinetochore, required for the microtubule attachment and chromatids separation (Yadav *et al.* 2018).

The centromeres of few budding yeasts have been characterised and found to be formed by short DNA sequences (< 400 bp). These are referred as “point” centromeres, and the most widely studied are in *S. cerevisiae* (a 125 bp sequence with three consensus centromeric DNA elements (CDEs), one of them non-conserved, (**Figure 1.6**) (Fitzgerald-Hayes, Clarke and Carbon 1982; Hieter *et al.* 1985) but they can also be found in other yeasts such as *C. glabrata* (153 bp) (Kitada *et al.* 1997), *K. lactis* (280 bp) (Heus *et al.* 1993), *Yarrowia lipolytica* (200 bp) (Fournier *et al.* 1993) or *C. maltosa* (325 bp) (Ohkuma *et al.* 1995).

Nevertheless, most organisms present longer centromeres, which are usually very repetitive and heterochromatic in nature. These are defined as “regional” centromeres and can also be further classified in small regional and large regional centromeres (**Figure 1.6**).

Small regional centromeres have been described in several *Candida* species, such as *C. albicans* or *C. dubliniensis* (Bensasson *et al.* 2008). These two yeasts present 8 chromosomes, and each one of the centromeres is formed by 3-5 kb sequences that have no common motifs or repeats, and therefore, each chromosome has a unique centromere sequence (Sanyal, Baum and Carbon 2004). The identification of putative centromeric sequences in other yeasts of the *Candida* clade, such as *C. lusitaniae*, *Scheffersomyces stipitis*, and *Debaryomyces hansenii* has proposed their classification as small regional centromeres, located in GC-poor regions and rich in retrotransposons in the case of *S. stipitis* and *D. hansenii* (Lynch *et al.* 2010).

The most studied large regional centromeres belong to the fission yeast *S. pombe*. These are located in a 10-15 kb CENP-A rich chromatin, flanked by 10-60 kb pericentric heterochromatin rich in repetitions, with a non-homologous unique sequence of 4-7 kb forming the central core (Steiner, Hahnenberger and Clarke 1993). Large regional centromeres have been also detected in filamentous fungi, such as *Neurospora crassa* (Centola and Carbon 1994) or *Aspergillus nidulans* (Aleksenko, Nielsen and Clutterbuck 2001) or in basidiomycetous such as *Cryptococcus neoformans* (Loftus *et al.* 2005), or *Coprinus cinereus* (Stajich *et al.* 2010).

1.2.3.4 Micro- and Minisatellites

The study of the genome of model organisms has allowed the identification of small repetitive regions, named micro- and minisatellites. Microsatellites are DNA regions of 20-60 bp length with tandem repetitions of 1-5 base pairs, being typical repeats units (GT)n, (CA)n, (CAA)n, (AT)n, or (GACA)n (Wöstemeyer and Kreibich 2002). On the other hand, minisatellites, consist of sequence motifs from 10-60 bp that are amplified to lengths of 0.1-30 kb.

These regions are quite ubiquitous, rapidly evolving and with relatively high rate of polymorphisms, but their position in the genome seems to be essentially stable (Wöstemeyer and Kreibich 2002). The polymorphisms are created by two different mechanisms (both in human and yeast cells), replication slippage, during the S-phase of the cell cycle, or repair slippage, which is associated with gene conversion.

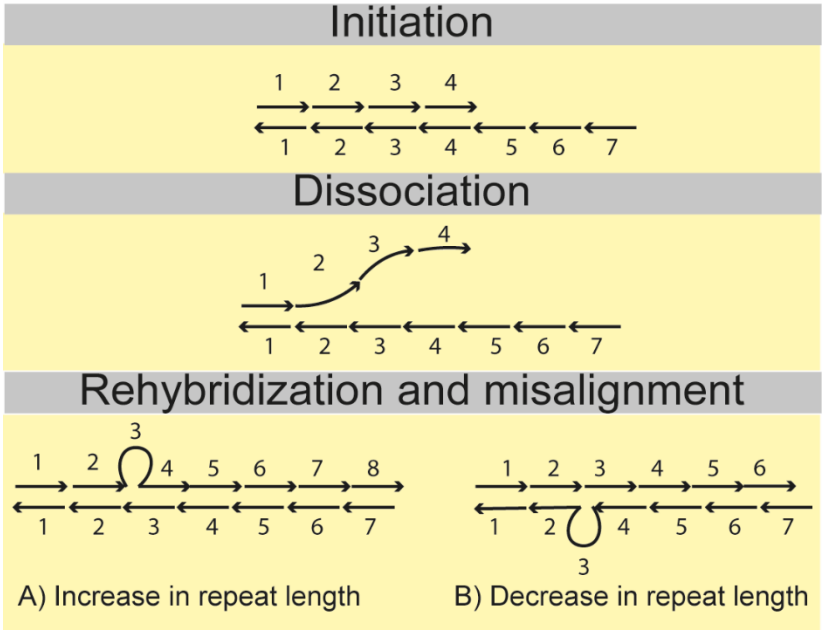


Figure 1.7. Replication slippage. After the replication of a repetitive region the two strands might dissociate, causing the pause of the DNA polymerase. If the strands then realign out of register, the number of repeats will change when the replication restarts.

During replication slippage, there is a transient dissociation of the replicating DNA strand, which is followed by a misaligned re-association. In this case, the DNA polymerase will pause in a way that only the terminal portion of the newly synthesised DNA is dissociated and separated from the template (Viguera 2001; Hile and Eckert 2004). This dissociated DNA strand will then realign to another repeat unit out of register, and the replication will restart, inserting or deleting repeats units relative to the template strand (Ellegren 2004) (**Figure 1.7**). Although the majority of these mutations will be corrected by the mismatch repair system, a small fraction will not be repaired and will

remain as a mutation (Strand *et al.* 1993). The only enzymatic activity required for this slippage is the DNA polymerase (Schlötterer and Tautz 1992).

The correlation between non-crossover gene conversion and tandem repeat rearrangements as polymorphism source (repair slippage) was described for the first time in meiotic expansion and contractions of the *CUP1* locus in *S. cerevisiae* (Welch, Maloney and Fogel 1990). This was subsequently demonstrated by Pâques *et al.* (Pâques, Leung and Haber 1998), when they induced an endonuclease DSB on a *S. cerevisiae* chromosome which could be then repaired using a homologous sequence containing tandem repeats as donor. Then, during the DNA synthesis associated with the DSB repair, the slippage of the newly synthesised strand may occur, causing an increase or decrease in the number of tandem repeats (Richard and Pâques 2000).

Despite of the early proposition of the relation of these polymorphisms with meiotic recombinational processes (Jeffreys *et al.* 1988), no positive correlation has been found between the location of minisatellites and the distribution of meiotic hot spots, at least in *S. cerevisiae*, and even when they are close to these hotspots they do not exhibit more polymorphisms. This seems to suggest that they are not acquired by mechanisms dependant of meiotic recombination (Richard and Dujon 2006).

Instead, the analysis of minisatellites in the genome of *S. cerevisiae* seems to indicate that they are located within genes that exhibit a negative GC skew (more cytosines than guanines) on the coding strand, and that they are even more skewed than the genes that contains them (Richard and Dujon 2006). Also their distribution seems to be associated with sites of recombination (Majewski 2000), although it is more likely that it is because of the nature of repetitive sequences being involved in the recombination process, rather than being the consequence of it (Treco and Arnheim 1986).

1.2.3.5 Formation of non-canonical DNA structures at repetitive loci

Repetitive loci can lead to the formation of abundant non-canonical DNA structures such as R-loops and G-quadruplexes that can be both beneficial or deleterious for the cells (Zhao *et al.* 2010).

R-loops are 3 stranded nucleic acid structures, formed by a RNA-DNA hybrid and a displaced ssDNA, that is often created during transcription, when the nascent transcribed transcript anneals to the DNA template strand (Hall *et al.* 2017) (**Figure 1.8, A**).

They have been shown to pause the progression of the replication fork, leading to DSBs, so they are considered as genome instability causing agents (Aguilera and

García-Muse 2012). Therefore, the prevention and resolution of R-loops is important and several proteins such as RNase H, topoisomerase enzymes and members of the THO complex (Hpr1, Thp2 and Tho2) are fundamental for this process (El Hage *et al.* 2010; Arora *et al.* 2014; Salvi *et al.* 2014).

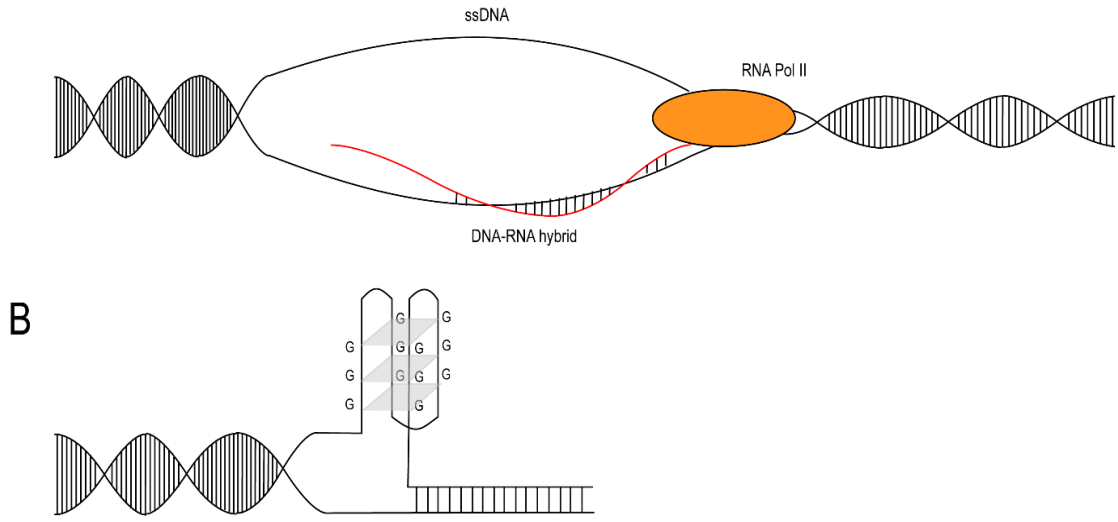


Figure 1.8. Non-canonical DNA structures commonly found in repetitive regions. (A) R-loops. (B) G-quadruplex with 3 G tetrads.

Classic examples of R-loop commonly found in yeast are the formed by a long non-coding telomeric repeat containing RNA (TERRA) in telomeres (Arora *et al.* 2014) or in the IGS regions of the rDNA locus(El Hage *et al.* 2010; Salvi *et al.* 2014).

G-quadruplexes (GQ) are tetrameric structures formed in DNA sequences with G4 $(GNxGNxGNx)_4$ motifs. GQs obtain their non-canonical square planar DNA structures (G-tetrads) by formation of Hoogsteen bonds between 4 guanine nucleotides with other 2 bordering guanines (Hall *et al.* 2017) (**Figure 1.8, B**). These loops vary in size (with smaller loops resulting in more stable structures) and can be classified according to: (i) Formation: Intramolecular or intermolecular when it is via one or multiple strands respectively (ii) Orientation of the constituent DNA: parallel or antiparallel (Hall *et al.* 2017).

G4 rich motifs with potential to form GQs are present throughout the genomes, and are particularly abundant in repetitive DNA loci, such as telomeric regions or the rDNA locus. They can also be formed on the displaced ssDNA of R-loops (Aguilera and García-Muse 2012).

Despite of their variable effects on genome stability, there are evidences of their role in telomeric regions, since they promote telomerase stability by promoting the binding of Pot and Ttp1, proteins of the shelterin complex (which are in charge of

recognising the telomeric repeats (de Lange 2005)) to the 3' ssDNA telomeric overhangs (Ray *et al.* 2014). They also act as telomere caps when the standard capping is compromised (Smith *et al.* 2011).

The presence of GQs can affect both directly and indirectly the stability of repetitive loci. The accumulation of GQs can impede the progression of the replication fork, thus impeding transcription, via GQ structures themselves or via stabilization of the ssDNA fragment of R-loops (Salvi *et al.* 2014; Paeschke, Capra and Zakian 2011). Also, triggering the replication fork may cause DSBs (Paeschke, Capra and Zakian 2011). To avoid this, yeast have developed a strategy to regulate their presence in transcribed regions: a DNA helicase (Pif1) binds and unwinds GQs maintaining the replication fork and promoting integrity (Paeschke, Capra and Zakian 2011; Ivessa, Zhou and Zakian 2000).

1.3 STRESS INDUCED GENOME INSTABILITY

Cells have different strategies to guarantee the stability of the genome and its replication and transmission to the offspring cells. Eukaryotes present checkpoints in the cell cycle that coordinate the synthesis of DNA with cell division, mechanisms for error detection and also DNA-damage repair pathways. However, sporadic replication or repair errors, or external genotoxic stresses can lead to genome instability (Aguilera and García-Muse 2013).

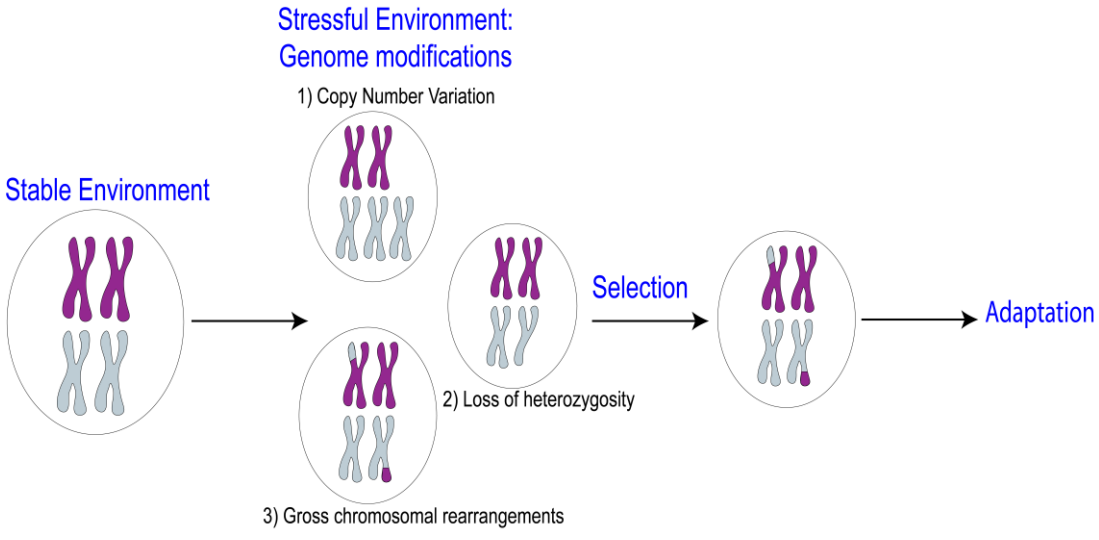


Figure 1.9. Common genome-wide modifications as consequence of genome instability derived of stressful environments. Populations under stressful conditions will suffer genome modifications, such as copy number variations, loss of heterozygosity or gross chromosomal rearrangements. The population most suitable to the new conditions will be selected and adapted.

Genome instability can result in alterations at different levels, such as mutations, variations in the chromosome number (or chromosome instability) or other complex

genetic alterations, like gross chromosomal rearrangements (GCRs), copy number variants (CNVs), hyper-recombination, or loss of heterozygosity (LOH) (Aguilera and García-Muse 2013) (**Figure 1.9**).

Mutations, which expand from point mutations to microsatellite expansions and contractions, are normally caused by erroneous, or error-prone, DNA synthesis, defective nucleotide/base excision repair (NER/BER), or mismatch repair (MMR). On the other hand, variations in the chromosome number is related to failure in the segregation machinery, or in the mitotic checkpoint (Aguilera and Gómez-González 2008). However, all these genomic alterations are in most cases initiated by single stranded DNA gaps, or double stranded DNA breaks (DSBs), generated as a consequence of replication stress, and will be repaired by different mechanisms, such as homologous recombination (HR), break-induced replication (BIR) and NHEJ (Aguilera and García-Muse 2013)

Under natural conditions, DNA is most vulnerable during replication in the S phase of the cell cycle. DNA synthesis machinery needs to overcome certain obstacles, such as secondary structure, or the presence of DNA-binding proteins, which can cause the stalling of the replication fork (RF), and therefore compromise the integrity of the genome (Aguilera and Gómez-González 2008). To avoid this, eukaryotic cells have developed systems that monitor the integrity of the DNA and coordinate this evaluation with DNA repair pathways, chromosome segregation and cell-cycle continuation. These checkpoints are important for the maintenance of genome stability and are highly conserved in eukaryotes (Myung, Datta and Kolodner 2001; Myung and Kolodner 2002).

When the obstacles in DNA during replication cause RF stalling, three main events can occur: (i) The RF remains associated to the replisome, and when the obstacle is removed the replication is resumed, (ii) The RF is delayed to allow better coordination between the fork processing and repair with replication continuation, or (iii) The RF collapses, which will cause the disassembly of the replisome and will create ssDNA gaps and DNA breaks (Aguilera and Gómez-González 2008; Sogo 2002; Cobb 2005).

These DNA breaks associated to replication can be generated in different ways. First, a RF can encounter a ssDNA nick, which will discontinue the DNA synthesis and cause a DSB. In case the nick is on the leading strand, the DSB would be one ended, and the re-start of the synthesis can be promoted via BIR. Secondly, the progression of the RF during the synthesis of the leading strand can be blocked, which will cause an uncoupling in the synthesis of the leading and lagging strands. This will promote the formation of a Holliday Junction (HJ) by the RF, which destabilises the replisome. The HJ can be reversed and the replication will go on, or it can be removed, and the synthesis will go on via BIR. Third, the synthesis of only one of the DNA strands can be blocked,

without the arresting the progression of the RF. For example, a lesion between two Okazaki fragments will not block the activity of the RF, but it will cause a gap in the lagging-strand. These possible errors can then be repaired by error-prone translation synthesis (TLS) or HR (Aguilera and Gómez-González 2008).

1.3.1 Understanding genome instability

Different studies have been conducted to identify the different mechanisms that affect genome instability in the cell. A commonly used approach among researchers has been the identification of agents that both prevent and promote genome instability. Three main contributors have been described: suppressors, fragile sites, and transcription.

Suppressors of genome instability are all those genes and proteins that act in trans to preserve genome integrity (Aguilera and Gómez-González 2008). Several studies in which different mutants suffered an increase in recombination events and chromosomal exchange have been useful to determine suppressors involved in replication (such as Sgs1 helicase (Gangloff *et al.* 1994), Rad27 Flap endonuclease (Tishkoff *et al.* 1997), or the nucleosome assembly factor Asf1 (Prado, Cortés-Ledesma and Aguilera 2004)), in the S-phase checkpoint (such as *Rfc5, Rad24* or *Mec1*) (Myung, Datta and Kolodner 2001; Myung and Kolodner 2002), or even in DNA-repair pathways (Myung, Chen and Kolodner 2001). However, not all GCRs arise as a defect in replication, since some are generated by telomere fusions (probably followed by breakage-bridge fusion cycles) (Mieczkowski *et al.* 2003), or by failures in the cohesion factors that promote equal segregation of chromatids exchange (De Piccoli *et al.* 2006).

Fragile sites are DNA regions that are commonly linked to hotspots for translocations, gene amplifications, integrations of exogenous DNA, or DNA rearrangements (Durkin and Glover 2007). They are normally ubiquitous DNA sequences that show gaps following partial inhibition of the DNA synthesis and can be classified in two types: common (95% of all known fragile sites), and rare (5%) (Aguilera and Gómez-González 2008).

Fragile sites are usually associated with trinucleotide repeats of the type CGG-CCG, CAG-CTG, GAA-TTC and GCN-NGC, with specific G-rich repeats, with DNA regions containing multiple tRNA genes, or with long A-T repeats (Mirkin 2007), since these repeats seem to be associated with the formation of secondary structures, such as hairpins or DNA triplexes, that commonly lead to instability related to perturbations during replication.

Certain DNA regions can suffer different modification during transcription, such as mutations (transcription associated mutations, or TAM), or homologous recombination (transcription-associated recombination, or TAR) (Kim and Jinks-

Robertson 2012). These modifications are strongly related with the block of the replication fork, by both the presence of RNA polymerase II in highly transcribed regions and the formation of secondary structures.

The formation of secondary structures during transcription is possible owing to the presence of transient ssDNA when the DNA double strand is opened. This ssDNA is chemically more unstable than dsDNA, so the nascent mRNA extruded from the RNA polymerase might interact with the transcribed strand, generating R loops, which can lead to genome instability, and therefore increase the TAR frequency. Moreover, the accumulation of transient ssDNA can lead to the formation of secondary DNA structures in repetitive segments, which might obstruct the action of the RF, leading to possible genome instability (Aguilera and Gómez-González 2008) (See section 1.2.3.5).

Negative effects of on the progress of RF have also been observed when transcription and replication are simultaneous, since a collision between the RNA polymerase II and the RF can lead to genome instability events that trigger TAR (Prado and Aguilera 2005).

1.3.2 Changes in chromosome number

Although most eukaryotic cells are euploid (their chromosomes have the same copy number) variations in the copy number of chromosomes is common in yeasts, and provide an important additional source of genome diversity (Comai 2005).

Aneuploid cells are those in which the copy number of one or more chromosome differs of those remaining in the genome. The maintenance of this different number of chromosomes in the population is an indication of the importance of the chromosomal copy number variation (CCNV) for physiological diversity (Gorter de Vries, Pronk and Daran 2017).

CCNV are caused by chromosomal missegregation during the anaphase of the cell cycle. Any error during this phase in any of the steps of the chromatid cohesion, centrosome formation at opposite cell poles, attachment of the microtubules to the kinetochore or the assembly checkpoints can be determining (Gorter de Vries, Pronk and Daran 2017; Thompson, Bakhoum and Compton 2010). In yeasts, this phenomenon can occur in both mitosis (Zhu *et al.* 2014) and, with a higher chance, meiosis (Parry and Cox 1970).

However, there are different factors that cause an increase in chromosome missegregation, such as chemical or physical stress factors (nutrient limitation (Adams *et al.* 1992), heat shock (Chen, Rubinstein and Li 2012) or irradiation with UV or X-ray (Parry *et al.* 1979)). Chemical stresses are also an important factor causing CCNV in

yeasts, such as fluconazole exposition in *Candida albicans* (Harrison *et al.* 2014), nocodazole in *S. cerevisiae* (Liu *et al.* 1997), and even high concentrations of ethanol have been reported to cause missegregation of chromosomes in fungal cells (Crebelli *et al.* 1989). Moreover, there are also genetic factors affecting chromosome missegregation, since aneuploid cells are more prone to acquire further CCNV (J. M. Sheltzer *et al.* 2011).

Once there is a missegregation and a cell becomes aneuploid, it will directly impact in their phenotype compared to the euploid cell. Typically, aneuploid yeasts show reduced fitness compare to euploids (Torres *et al.* 2007). Firstly, aneuploidy increases the genome instability in cells, which might be related with problems in further chromosome segregation and mitotic recombinationb(Blank *et al.* 2015; Skoneczna, Kaniak and Skoneczny 2015; Gorter de Vries, Pronk and Daran 2017). Secondly, there are changes in the gene expression, which normally involve an increase in the regulation of genes related to the response to environmental stresses, and a decrease in the genes involved in cell growth and proliferation (Torres *et al.* 2007; Sheltzer *et al.* 2012). Moreover, the genes linked to cell wall metabolism are also normally affected, which might be important for the adjustment of the cell wall with the increase in size that can be related to aneuploidy (Wu *et al.* 2010).

Moreover, the genes that are carried by the gained or lost chromosome also change their expression. In case of chromosome gain, there will be an overexpression, that can lead to the accumulation of un- or misfolded proteins, which might cause proteotoxic stress (Oromendia, Dodgson and Amon 2012). The extra energy required for the processing of the over produced proteins seems to be related with the increased nutrient consumption and slower growth of aneuploid yeasts (K. A. Geiler-Samerotte *et al.* 2011).

On the other hand, beneficial effects in cellular fitness have also been described. First, CCNV offers new copies of genes, which will allow a fast way increase evolution by neofunctionalization (Rancati *et al.* 2008; Rancati and Pavelka 2013). Second, under certain selective conditions, the aneuploid strain might outgrow the parental population, and any further mutation will enhance positive effects of the CCNV (Gorter de Vries, Pronk and Daran 2017).

1.3.3 Changes in chromosome structure

Changes in chromosome structure are called rearrangements, and can be classified in mainly four events: deletions, duplications, inversions or translocations, in which chromosome segments are lost, doubled, orientation-inverted or moved to another location, respectively (Griffiths 2012).

The generation of chromosome rearrangements requires at least two simultaneous phenomena. First, the breakage of the chromosomes in at least two different points is necessary, and second, the DNA repair machinery needs to repair this damage joining the ends of two different breaks in a way that the new chromosome structure generated presents one centromere and two telomeres. If the new chromosome structure lacks the centromere (acentric) it will not be able to join to the microtubules during the anaphase and therefore will not be present in the progeny of the cells. In the case that telomeres are not present, replication can be compromised (Griffiths 2012).

Moreover, rearrangements as a result of crossing overs between repetitive DNA segments in the same, or different chromosomes, have also been described. This phenomenon is called non allelic homologous recombination (NAHR) (Griffiths 2012).

The changes in the chromosome structure can be balanced or unbalanced. Balanced rearrangements change the order of the genes in the chromosomes, but do not remove or duplicate any section of DNA, such is the case for inversions and reciprocal translocations. On the other hand, unbalanced rearrangements will change the gene dosage in the chromosomes, with increases in the case of duplications (Griffiths 2012).

Deletions can be intergenic, in which a part of the gene will be removed, and therefore the gene inactivated, or multigenic. On the other hand, duplications can be located adjacent to each other (tandem duplications) or somewhere else in the genome (insertional duplications). Examples of all these phenomena have been described in yeasts (Yang, Ohnuki and Ohya 2014; Kishida *et al.* 1996; Achaz *et al.* 2000; Koszul, Dujon and Fischer 2006).

Inversions require the break of a chromosome segment in two points, and the reinsertion of the segment in the opposite orientation. When the centromere is part of the inversion it is known as pericentric inversion, whereas when it is not affected the inversion is paracentric. Finally, in translocations, two chromosomes trade acentric fragments created by two simultaneous chromosome breaks. Both have also been widely reported in yeasts (Seoighe *et al.* 2000; Naseeb and Delneri 2012; Lemoine *et al.* 2005; Shibata *et al.* 2009).

These structural changes in yeast can lead to modification in phenotypes. Morphological switches have been described in yeasts like *C. albicans* through homologous recombination dependant chromosomal translocations (Andaluz, Ciudad and Larriba 2002; Legrand *et al.* 2007), and the deregulation of the actin network and

over transcription of ABC transporters have also been reported in yeasts carrying translocations (Nikitin *et al.* 2014).

After the chromosomal translocation, cells undergo and adaptation, following an arrest of the cycle at G1 or G2/M (Nikitin *et al.* 2008). Cell carrying translocations are highly unstable and are prone to further rearrangements or even aneuploidies (Tosato, Sidari and Bruschi 2013).

Nevertheless, some chromosome translocations can be a benefit, since they can act as genetic basis of variability for complex molecular adaptation to changes in environmental conditions (Tosato and Bruschi 2015). Positive effects in fitness have been reported for both reciprocal (Colson, Delneri and Oliver 2004) and non-reciprocal translocations (Nikitin *et al.* 2014) in yeasts. For example, they can increase resistance to drugs used against human pathogens (Ahmad *et al.* 2013), or suffer proteomic changes that allow industrial fermentations at different conditions (Paget, Schwartz and Delneri 2014; García-Ríos, López-Malo and Guillamón 2014).

1.3.4 Loss of heterozygosity

Diploid cells often suffer partial losses of regions of the genome with respect to their parent. This phenomenon is known as loss of heterozygosity (LOH), and it generates genetic variability by exposing phenotypes that are related to recessive alleles (Forche *et al.* 2011). The regions lost can range from short tracts, via gene conversion or double crossovers, to long tracts that arise via single crossover events or by non-reciprocal events, such as BIRs, where a dsDNA break will be repaired by invasion of the broken end into a homologous DNA sequence (Freire-Benéitez, Gourlay, *et al.* 2016; Forche *et al.* 2011). Whole-chromosome LOH events have also been described, and they eventually lead to aneuploidy (Forche *et al.* 2011).

Normally, the two alleles in the new heterozygous strain will provide a differential phenotype, which could be beneficial, deleterious or detrimental depending on the environmental conditions, which highlights the importance of their study in yeast adaptation and evolution experiments.

1.3.5 Studying changes in the genome caused by instability

Several studies in yeasts like *S. cerevisiae* (Argueso *et al.* 2008; Paek *et al.* 2009; Tang *et al.* 2011) or *C. albicans* (Freire-Benéitez, Price, *et al.* 2016; Robert T Todd *et al.* 2019) have demonstrated that the way genome modifications affect different loci are often related to the features of the local DNA, such as proximity to repeats or transposons, can lead to chromosome rearrangements, chromosome aberrations or general genome instability (Forche *et al.* 2011). Hence, it is important to develop

techniques that allow the study of genome modifications and the characteristics of the DNA affected.

The analysis of genome modifications rely on several techniques that are complementary: Array comparative genomic hybridization (aCGH), Contour-clamped homogeneous electric field (CEHF) electrophoresis, flow cytometry, fluctuation analysis, fluorescence in situ hybridisation (FISH), quantitative real-time PCR (qPCR), or next generation sequencing (NGS) (Gorter de Vries, Pronk and Daran 2017) are widely used.

Plenty of publications have focused on the study of LOH in several yeasts. One of the most widely used techniques is the fluctuation analysis (Luria and Delbrück 1943). This assay is based on the measurement of how different events fluctuate from one culture to another, and relies normally on the study of the loss of different genetic markers that condition for selectable phenotypes, whose distribution in a population will serve as an estimation of spontaneous mutations (Spell and Jinks-Robertson 2004). This technique has been established for *S. cerevisiae* (Andersen *et al.* 2008), but has been widely used also for the determination of both LOH (Forche *et al.* 2011) and aneuploidy (Bouchonville *et al.* 2009) in *C. albicans*.

The detection of chromosome rearrangements is possible with CHEF electrophoresis. This technique separates big fragments of DNA (up to 9 megabases (Cox *et al.* 1990)) in an agarose gel. This resolution allows the separation of chromosomes, and therefore analyse the karyotypes of different strains (Chu, Vollrath and Davis 1986; Török, Rockhold and King 1993), which therefore, allows also the determination of CCNV. Combination of CHEF with Southern hybridization allows the determination of individual chromosomes, specific translocations, or even duplications by comparison of the hybridization intensity between strains (Waghmare and Bruschi 2005).

Chromosome rearrangements can also be detected by both FISH and aCGH. FISH uses probes labelled with fluorophores to locate the specific positions of DNA sequences on chromosomes. Therefore, it is easy to determine also copy number variations. These rearrangements and copy number variations can be also detected by aCGH. This technique is based on the use of a fluorophore to mark the genomic DNA of interest. Then this extracted DNA will be hybridised in microarrays that have small segments of target DNA, increasing the resolution of the technique (Theisen, A 2008).

Copy number variations can be detected by quantitative real-time PCR (qPCR), which is based on the use of primers that amplify specific regions in the genome (Pavelka *et al.* 2010). However, this amplification will be specific for small regions of the chromosomes, and other validation methods are required for the determination of a

complete CCNV. One of these techniques is flow cytometry, which allows the analysis of DNA content and ploidy using fluorescent DNA-intercalating dyes. Moreover, when the dye used does not compromise the viability of the cells, these can be sorted according to DNA content (Fluorescence-activated cell sorting, or FACS) even when this differs less than 2% (Pfosser *et al.* 1995), so this population can be then further studied. This high resolution highlights this technique as a very powerful tool for the study of CCNV.

Finally, next generation sequencing (NGS) techniques for analysis of whole genomes offer high resolution and accurate analysis of genome modifications (Xie and Tammi 2009). Several techniques, such as Illumina (Heather and Chain 2016), SOLid (Shendure *et al.* 2005) or Ion Torrent (Rothberg *et al.* 2011) have been widely used for the determination of genome sequences. However, despite of being extremely powerful, there are some drawbacks in the use of NGS technologies. One of the main problems is that these technologies are based in the use of relatively short reads, which leads to misassemblies and gaps in repetitive regions of the genome, when the repetition is larger than the read size. Moreover, the determination of large structural variations with small reads is also complicated. Furthermore, they rely on PCR, which causes difficulties in regions with high GC content, which are inefficiently amplified (van Dijk *et al.* 2018).

These problems have recently been overcome with the use of third generation sequencing (TGS) techniques, such as single-nucleotide real time (SMRT) sequencing (Eid *et al.* 2009), released by Pacific Biosciences (PacBio) in 2011, or Oxford Nanopore Technologies (ONT), released in 2014 (Jain *et al.* 2015). Both technologies are based on the sequencing of large fragments of DNA, which allows the assembly of repetitive regions, such as rDNA locus, telomeres, or subtelomeric regions, offering more accurate depth analysis (van Dijk *et al.* 2018).

The great potential of these long-reads sequencing techniques lies in the possibility to capture entire chromosome modifications in a single read, which will unravel chromosome structures, translocations breakpoints, and duplications of chromosomes or its fragments (Istace *et al.* 2016). Therefore, the identification of chromosome rearrangements is now possible through alignments of strains with reference genomes, and duplications and CCNVs can also be detected by several methods, for example, via read depth by mapping the sequencing reads versus a pre-assembled genome or a *de novo* assembly genome (Zhao *et al.* 2013), or via the determination of the allele frequency in both the whole genome or in specific regions (Yuan O Zhu, Sherlock and Petrov 2016).

Nevertheless, TGS techniques are less accurate than past technologies (Eid *et al.* 2009; Schneider and Dekker 2012), and hence, they require sensitive alignment methods and a limit in the discrimination of divergent alleles and non-exact repeats (Koren *et al.* 2017). Moreover, the assembly programs are normally designed for short and highly accurate reads, so they have been challenged by the increased read length and error rate of TGS technologies.

To solve these problems, different approaches have been addressed in the last years, and have been classified as hybrid, hierarchical or direct methods. Hybrid methods use single molecule reads to reconstruct the genome, but they rely on complementary short reads obtained with other technologies for accurate base calls. Hierarchical methods use multiple rounds of alignments and corrections to improve the quality of the reads prior to the assembly, but it does not require a secondary technology. Finally, direct methods attempt to assemble the reads from a single overlapping step without any read correction.

1.4 GENOME INSTABILITY EFFECTS ON INDUSTRIAL STRAINS: USING YEASTS FOR BIOETHANOL PRODUCTION

The use of yeasts in biotechnology is very extensive due to their high number of properties and the fast development of methodologies and applications. During the past decade, different yeast systems have been established to produce heterologous proteins with post-translational modifications similar to those observed in humans, but also other high value chemicals and proteins can be produced. Moreover, their use for environmental applications has gathered attention, since they can be used for bioremediation processes or production of biofuels due to their metabolic capabilities (Johnson and Echavarri-Erasun 2011).

In the last years, research on the use of genetically engineered strains to produce compounds of interest is increasing, since improvements are normally required for economic viability of the processes (Mans, Daran and Pronk 2018). However, these modifications require specific knowledge on the pathways to be altered, and the effects observed are not always anticipated.

Evolutionary engineering, also known as adaptive laboratory evolution (ALE) is a complementary strain improvement strategy that exploits the genome plasticity of microorganisms by applying prolonged selective pressures during their cultivation (Dragosits and Mattanovich 2013). This will allow the isolation of adapted strains that

can be further analysed for phenotypic or genotypic changes when compared with their parental. Moreover, the whole genome of these strains can be subsequently sequenced to try to understand which changes in the genome are responsible for those beneficial traits (Liu and Zhang 2019) (**Figure 1.10**).

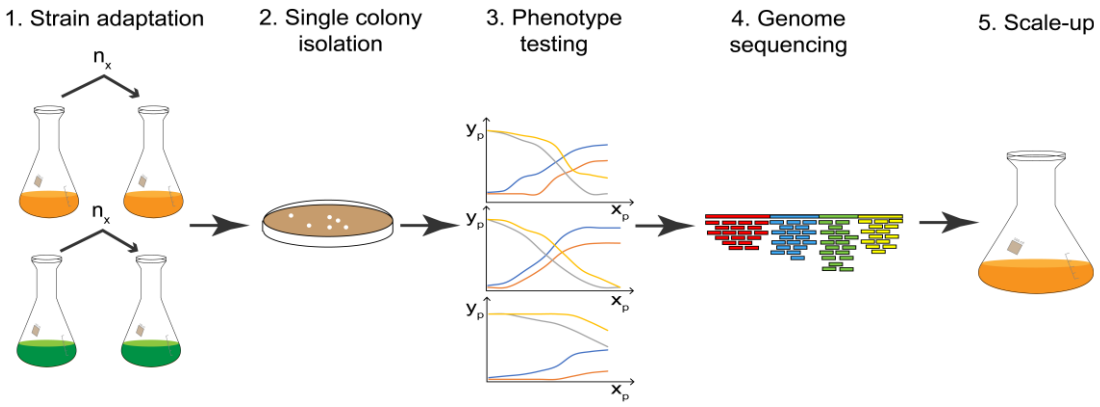


Figure 1.10. Adaptative laboratory evolution (ALE) strategy. Strains adapted to different conditions during N generations are isolated as single colonies and classified according to desirable traits. The top strains can then be sequenced and if the genes responsible of the improved traits are detected, genetic engineer and scaling-up approaches can be used for further improvements.

However, ALE strategies are based on the idea that environmental stress conditions induce genome modifications, which has been controversial because of a lack of distinction between the stress conditions themselves, and the stress conditions required to select mutations. However, this idea is widely accepted nowadays, and *S. cerevisiae* is currently a well-established organism to study genome instability (Wang 2008).

This approach has been widely used for the isolation of strains with CCNV linked to improvements in industrial phenotypes, such as higher resistance to inhibitors (G. Chen *et al.* 2012), products (Voordeckers *et al.* 2015), improved kinetics in fermentation (Sato *et al.* 2014) or in sedimentation (Oud *et al.* 2013). Several of these studies have been conducted to improve strains used for bioethanol fermentation.

Bioethanol, also known as EtOH, ethyl alcohol, or chemically C_2H_5OH (Mohd Azhar *et al.* 2017), is one of the most promising alternative liquid fuels to petrol. The term is only used to designate the amount of ethanol that is going to be used as fuel (Parapouli *et al.* 2020). Despite it has a 68% lower energy compared to petrol, the high oxygen content in ethanol makes its combustion cleaner, which results in lower emission of toxic substances (Aditiya *et al.* 2016). Moreover, it is biodegradable, has broader flammability, higher octane number, increased heat of vaporization and higher flame speed (Mohd Azhar *et al.* 2017).

Depending on the feedstock used for their production they can be classified as first, second and third generation bioethanol. First generation bioethanol involves the use of biomass rich in starch (corn, wheat, rice, potato, cassava or barley) or sucrose (sugar cane, sugar beet, or fruits). Second generation bioethanol is produced from lignocellulosic biomass, such as wood, straw and grasses. Finally, third generation bioethanol is derived from micro- and macroalgae biomass (Nigam and Singh 2011). From these, second generation bioethanol, obtained from lignocellulose is one of the most promising, mainly because of their great availability, low cost, and non-competence with the food demand (Hemansi *et al.* 2019).

Lignocellulosic biomass is mainly composed of cellulose, hemicellulose and lignin. Cellulose represents 40-60% in weight, and it is formed by fibrils of cellobiose, which is formed of polymers of glucose bound together by glycosidic bonds. Hemicellulose accounts for 25-35% of the total weight of lignocellulose, and it is formed by polymers of different sugars, such as xylans, xyloglucan, mannans and glucomannans forming an amorphous protective matrix. Lignin is a tri-dimensional heteropolymer of phenylpropanoid units (p-coumaryl, coniferyl and sinapyl alcohol), which represents 15-40% of the total weight. It binds hemicellulose and cellulose in the cell wall, and it is the main responsible of the hydrophobicity and rigidity of lignocellulose. Moreover, there are other minor components, such as proteins, terpenic oils, fatty acids, and inorganic materials (Lange 2007; Zoghlami and Paës 2019).

The complexity of this materials make lignocellulose a very recalcitrant material, so different methods (chemical or physical) are required for the conversion of polysaccharides into fermentable sugars (Zoghlami and Paës 2019).

The production of ethanol is based on the ability of microorganisms to catabolise carbohydrates into two carbon components. This process begins with the conversion of six carbon sugars (glucose) into pyruvate in a 10 reaction metabolic pathway called glycolysis. Then, in the presence of oxygen pyruvate enters the tricarboxylic acid (TCA) cycle and proceeds to the respiration chain to produce ATP and carbon dioxide. On the other hand, in the absence of oxygen, pyruvic acid can follow two routes, depending on the cell: conversion to ethanol and carbon dioxide (alcoholic fermentation), or conversion to lactate (lactic acid fermentation) (Alba-Lois, L. and Segal-Kischinevzky, C. 2010).

Therefore, alcoholic fermentation pathway will require the detention of pyruvate from entering the TCA cycle and, subsequently, the inactivation of the respiratory chain. The mechanisms to avoid it will depend on the yeast physiology. In Crabtree positive yeasts, such as *S. cerevisiae*, the respiration pathway is avoided in the presence of high concentration of sugars, so two carbon sugars metabolism is promoted, and therefore

ethanol is accumulated (De Deken 1966). However, after their depletion, the catabolism will shift from two carbon sugars to CO_2, phenomenon known as diauxic shift (DeRisi 1997). On the other hand, in Crabtree negative yeasts, such as *C. tropicalis,* or *S. stipitis*, the sugars are always catabolised to CO_2, and therefore the respiratory pathway can only be avoided in the absence of oxygen (De Deken 1966; Prior BA, Kilian SG and Du Preez JC 1989).

The selection of the different yeasts strains to be used for fermentation will depend on their characteristics, but the ability to resist stressful conditions, such as increases in ethanol concentration (over 20%) and temperature (35-45 ºC) (Banat *et al.* 1998), osmotic stress (D'Amore *et al.* 1988), presence of growth or fermentation inhibitors (Amin, Standaert and Verachtert 1984), or contamination by other microorganisms (Beckner, Ivey and Phister 2011) are fundamental in any industrial fermentation process.

Nevertheless, there are several challenges that any selected microorganism will face in these processes. During fermentation, biomass proliferation and the increase in the metabolism will cause a rise in the temperature, which might inhibit not only the growth and viability of microorganisms, but also ethanol production (Nicolaou, Gaida and Papoutsakis 2010). To avoid this, use of thermo-tolerant strains is recommended, which will also reduce processing costs, since the fermentation medias will not require cooling steps (Banat *et al.* 1998). Strains adapted to high concentrations of inhibitors and ethanol have also been developed (Dinh *et al.* 2008; Landaeta *et al.* 2013; Narayanan *et al.* 2016). Moreover, flocculent strains have been widely used, which allows the separation of the biomass and fermentation media without centrifugation, reducing the downstream processing costs and permits operation at higher cell density, achieving then higher overall productivity (Soares 2011).

S. cerevisiae is the most used yeast in industrial ethanol production, especially in the wine and brewery industries. The extensive knowledge in its genetics, physiology and methodology facilitates its applicability. Moreover, it has both high ethanol yield and productivity, resistance to high ethanol concentration and other stresses, and high adaptation for scale up processes (Parapouli *et al.* 2020). However, one of the main problems of the use of *S. cerevisiae* in ethanol production is its inability to ferment pentose sugars (Parapouli *et al.* 2020).

This can be solved using hybrids (hybrid strains of *S. cerevisiae* and other pentose fermenting yeasts, such as *Pachysolen tannophilus*, *Candida shateae* or *S. stipitis* have been developed by protoplast fusion (Kumari and Pramanik 2012)), genetic engineered strains with genes required for pentose uptake and metabolism (Jin, Laplaza

and Jeffries 2004; Hou *et al.* 2017; Osiro *et al.* 2019), or by the co-culture of two yeast strains, one with the ability to ferment hexose, and the other pentose (Taniguchi *et al.* 1997; Unrean and Khajeeram 2015; Pathania, Sharma and Handa 2017).

Nevertheless, the results obtained with these approached are not optimal, and pentose sugars are not always consumed completely, or other by-products, such as xylitol are also formed (Toivari *et al.* 2004). This reduces the efficiency of the process, since not all the available sugars are converted to ethanol.

In the last years, different approaches are being combined to try to overcome the difficulties in the improvement of bioethanol industrial fermentation. Traditional genomics allow the determination of single nucleotide polymorphisms (SNPs) and quantitative trait loci that provide an explanation for improvements in phenotypic traits of interests, such as high ethanol tolerance (Swinnen *et al.* 2012; Haas *et al.* 2019), thermotolerance (Wang *et al.* 2019), or modification in the flux distribution of the carbon metabolism (Eder *et al.* 2020). Nevertheless, the use of next generation sequencing techniques combined with traditional genomics have allowed stablishing new pipelines to understand how genome modifications shape phenotypes (Salazar *et al.* 2017; Tiukova *et al.* 2019). PacBio and Nanopore have been useful for the determination of genome variation in ethanol producing yeasts, and an indication of genome plasticity in both industrial and laboratory strains (McIlwain *et al.* 2016; Vassiliadis *et al.* 2020).

The combination of these approaches can be used to fill the gaps in the knowledge of pentose fermentation and in the genetics and physiology of pentose fermenting yeasts, and hence, target the genome modifications that need to be conducted for more efficient ethanol fermentation.

1.5 ETHANOL FERMENTATION FROM PENTOSES: THE CASE OF *SCHEFFERSOMYCES STIPITIS*

The first screening of yeasts that were able to ferment pentoses (xylose) was conducted by Toivola *et al*. (Toivola *et al.* 1984), and involved more than 200 yeasts.

Phylogenetically, all xylose-fermenting yeasts studied belong to the CTG clade (Wohlbach *et al.* 2011), formed by *Candida* species, *S. stipitis, Debaryomyces hansenii* or *Lodderomyces elongisporus*. These species present a modification from the universal genetic code, translating CUG codons in the mRNA by serine (ser) instead of leucine (leu) (Santos and Tuite 1995). The change of identity in the codon is mediated by a $tRNA_{CAG}^{Ser}$, which is able to synthesise two types of aminoacyl tRNAs: Ser-$tRNA_{CAG}^{-Ser}$

and Leu-$tRNA_{CAG}^{-Ser}$, which will compete for the recognition of CUGs during translation (Santos *et al.* 2011). Comparative genomics studies showed that the a $tRNA_{CAG}^{Ser}$ appeared 272±25 million years ago, before the divergence between *Saccharomyces* and *Candida* genera, via insertion of an adenosine in the anticodon of the serine $tRNA_{CAG}^{Ser}$ gene (Massey *et al.* 2003). However, the evolutionary pathway of the CUG reassignment is not completely understood.

Fermentation studies conducted by du Preez and Prior (du Preez and Prior 1985) have shown that *S. stipitis* is the yeast with most ethanol productivity from xylose from all the known xylose-fermenting yeasts. *S. stipitis* (previously named *Pichia stipitis*) was first isolated from insect larvae by Marie-Claire Pignal in 1967(Pignal 1967), and several studies have demonstrated an endosymbiont relation with passalid beetles, which inhabit and degrade white-rooted hardwood (Suh *et al.* 2003; Urbina, Schuster and Blackwell 2013)(**Figure 1.11**).

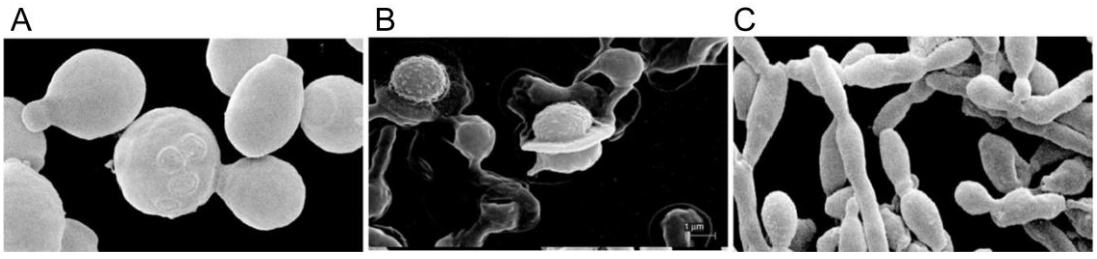


Figure 1.11. *Scheffersomyces stipitis* growing exponentially in buds with scars (A). *S. stipitis* spores with hat-shape seen from top and side (B). Pseudomycelia formed under carbon-limited continuous culture (C). Images obtained from Jeffries *et al.* (2007)

Phylogenetically, *S. stipitis* is a hemiascomycetous (Saccharomycetes) budding yeast, and as such it is included in the division Ascomycota. It belongs to the order Saccharomycetales and the family *Saccharomycetaceae* (Jeffries *et al.* 2007).

During vegetative growth, *S. stipitis* forms buds along pseudomycelia (Jeffries *et al.* 2007), and it has an haplontic life cycle, which means that it is mostly haploid, although it is able to undergo conjugation, and then almost immediately goes through meiosis (Melake, Passoth and Klinner 1996) through the formation of two hat-shaped ascospores from each ascus (Jeffries *et al.* 2007). The organization of the MTL locus suggests that it is an homotallic yeast, which means that this conjugation can only occur between genetically identical cells. The single MTL locus presents genes from both MTLa and MTLα (a1, a2 and α1), but there are also other non-sex genes associated to it, such as *OBP* (oxysterol binding protein gene), *PIK* (phosphatidylinositol kinase gene) and *PAP* (poly(A) polymerase gene), resembling to the MTLα of other species (Butler 2010). The presence of other regulatory genes, such as *GAP1* (General amino acid permease) or *CBP1* (Corticosteroid binding protein gene) has been confirmed (Wolfe and Butler

2017). This structure seems to have been originated from a recombination between ancestral MTLα and MTLa idiomorphs, placing the regulatory genes together (Butler 2010). Notably, the α2 gene is absent, which correlates with other sexual species of the CTG clade, and explains differences in sporulation between them and Saccharomyces species (Butler *et al.* 2009; Butler 2010) (**Figure 1.12**).

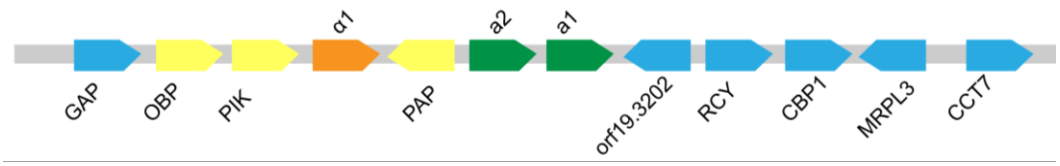


Figure 1.12. Mating type (MAT) locus in *Scheffersomyces stipitis*. Blue boxes indicate genes with regulatory effects over sex-related genes. Yellow boxes indicate non-sex genes associated to the MAT locus.

1.5.1 The genome of *S. stipitis*

The genome of *S. stipitis* (strain NRRL Y-11545 = ATCC 58785 = CBS 6054 = IFO 10063) has been completely sequenced using a shotgun approach offering a high quality genome assembly (error rate lower than 1/100,000) (Jeffries *et al.* 2007). The complete genome has a size of 15.4 Mbp and is organised in 8 chromosomes, ranging from 3.5 Mbp to 0.97 Mbp, with one gap in chromosome 1 (Jeffries *et al.* 2007). Moreover, the whole genome of the strain NRRL Y-7124 has also been sequenced as part of a yeast genome evolution comparison (Shen *et al.* 2018), although to our knowledge no further analyses have been conducted about its structure.

The first annotation conducted by Jeffries *et al* (Jeffries *et al.* 2007) predicted a total of 5,841 genes, using the Joint Genome Institute (JGI) Annotation Pipeline, with the majority of them (72%) presenting a single exon. Moreover, a function can be predicted to about 70% of the genes according to KOG (eukaryotic orthologous groups) classifications. Approximately 84% of the proteins showed strong similarity to proteins present in other fungi, being *D. hansenii* the yeast that offered highest genome conservation by bidirectional BLAST analysis of the gene models. A phylogenetically inferred group comparison (PhIGs) identified 25 gene families, representing 72 proteins, specific to *S. stipitis*, whereas 81 gene families, representing 442 proteins, were missing, relative to other yeast genomes.

The most frequent protein domains characterised are protein kinases, helicases, transporters (for sugars and major facilitator superfamily (MFS)), and domains involved in transcriptional regulation. Although the majority of these domains are shared with other

hemiascomycota (1534 domains shared with *S. pombe* or 1639 with *D. hansenii*), specific domains of *S. stipitis* have been described, which belong to glycosyl hydrolase families, a subgroup of cellulases and xylanases (Jeffries *et al.* 2007).

Despite of the high accuracy of the genome sequence published by Jeffries *et al*. (Jeffries *et al.* 2007), Maguire *et al*. (Maguire *et al.* 2013) identified annotation problems during a comparative analysis of yeast genomes of *Candida* species and other yeast with biotechnological applications, including *S. stipitis*. In their publication they added 211 genes, for an updated total of 6,026. Moreover, they noticed that, in the original annotation, 1,611 genes included introns, with multiple introns in several genes (for a total of 2,567 introns), which seemed surprisingly high for species in the Saccharomycotina (*C. albicans* presents 415 introns in 381 genes (Mitrovich *et al.* 2007), and *C. parapsilosis* 422 in 387 genes (Guida *et al.* 2011)). Since none of the introns had experimental support they carried out a closer investigation on them to find out that they were in-frame, and when included in the translation, the predicted protein generated had greater similarity to orthologs in other species than the spliced gene model. Therefore, with this approach they removed 1,231 introns from the gene models, yielding to a total of 380 introns. Despite of this the annotation and gene identifiers of the original *S. stipitis* data base has been maintained (Maguire *et al.* 2013).

The genome presents 8 GC-poor troughs, one per chromosome, that seem to be associated with the centromeres. These regions are highly sequence-repetitive, and rich in clusters of Ty5-like LTR retrotransposons (Tps5), which appear in a mixture of intact and truncated elements, and with long terminal repeats associated to them (Lynch *et al.* 2010). Cao *et al*. (Cao, Seetharam, *et al.* 2017; Cao, Gao, *et al.* 2017) reported the discovery of a 125 bp region in the GC-poor troughs that was able to significantly increase plasmid stability, and reported them as point centromeres in *S. stipitis*. The lack of consideration of the Tps5 clusters in Cao's studies led Coughlan and Wolfe (Coughlan and Wolfe 2019) to re-examine their proposed centromeres to conclude that the 125 bp sequence proposed by Cao *et al* is actually part of a long terminal repeat of a Tps5, and identified three different Tps5 elements (Tps5a, Tps5b and Tps5c), with different long terminal repeats associated to them (LTRa-c) and a non-coding, non-autonomous large retrotransposon derivative (LARD) with two associated long terminal repeats in parallel orientation (LTRd) as forming elements of regional centromeres (**Figure 1.13**).

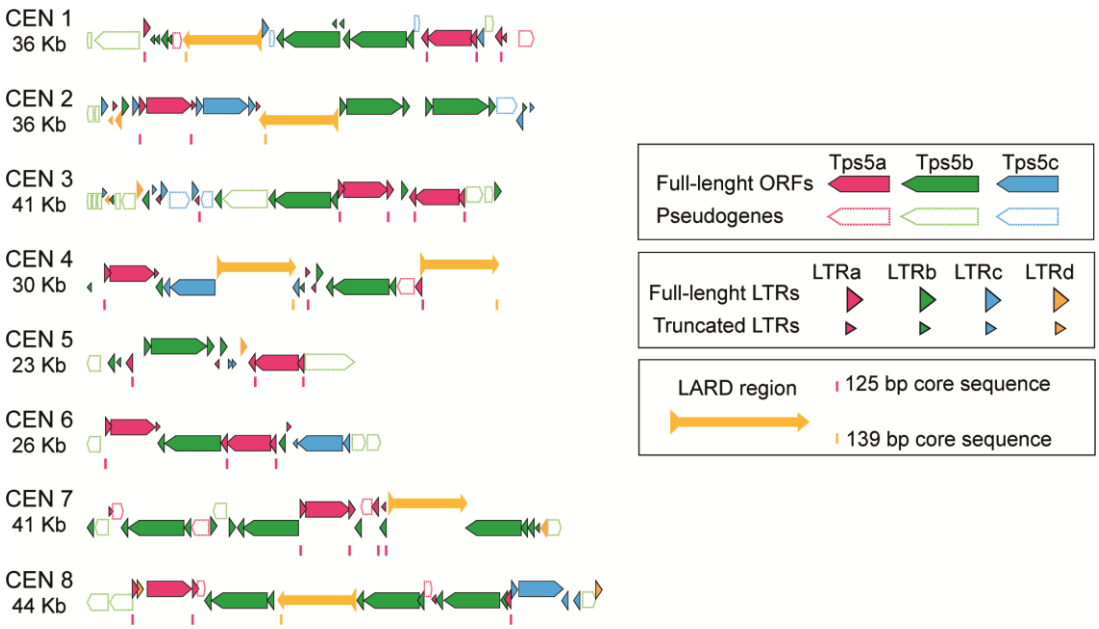


Figure 1.13. Regional centromere structure in *S. stipitis*. Adapted from Coughlan and Wolfe (2019). Abbreviations: CEN: centromere. LTR: Long terminal repeat. LARD: Large retrotransposon derivative. Tps5: Ty-5 like transposon.

1.5.2 Gene cluster organization

Apart from the highly repetitive sequences found at the centromeres, the genome of *S. stipitis* is also characterised by the presence of several tandem repeats and gene clusters. Tandem duplications are normally an early stage in the formation of gene families. The maintenance of genes resulted of a duplication event are an indication of both the feasibility to adapt to different niches and also of the high levels of enzymatic activity required for survival.

The most noticeable functional clusters in *S. stipitis* consist of two or more families of homologs associated in clusters. Jeffries and Van Vleet (Jeffries and Van Vleet 2009) defined gene families in *S stipitis* as proteins that had consistent phylogenetic relationships and common domain architectures. They identified a total of 35 gene clusters (**Table 1.1**), of which 5 appear in subtelomeric regions of chromosomes (<30Kbp from the end), where recombination occurs frequently. 18 clusters consist of two or more genes that have different enzymatic activities, but are functionally related, such as maltose permeases, found in association with α-glucosidases, or ß-glucosidases, found in association with endoglucanases. Notably, some of the gene clusters found in *S. stipitis* seem to be highly conserved across species, such as urea metabolism or histones.

The genes present in a gene cluster can be in a convergent, divergent, or tandem orientation, depending on the direction of the transcript. In divergent orientation both

Table 1.1. Functional gene clusters in *Scheffersomyces stipitis*. (*): Gene cluster is present in the subtelomeric regions. Conv: Convergent, Div: Divergent. Tand: Tandem

Cluster name	Genes in cluster	Location (Kb)	Orientation
α-Glucosidase	*MAL1-MAL6*	Chr2 752.5 - 757.2	Div.
α-Glucosidase	*MAL2-MAL7 (SUC1.1)*	Chr5: 4.5 - 9.3 *	Div.
α-Glucosidase	*MAL3-AGL1-YIC1-MAL5 (SUC1.4, SUC1.2)*	Chr6: 26 - 42 *	Div; Div
β-Galactosidase	*LAC1-LAC4*	Chr2: 1781.2 - 1786.9	Div.
β-Galactosidase	*BMS1-LAC3*	Chr3: 24 - 25	Div.
β-Glucosidases	*HXT2.4-EGC2-BGL5*	Chr1: 614 - 626	Div; tand.
β-Glucosidases	*BGL2-HCT2.3-HGT2*	Chr2: 2707.5 - 2721 *	Div; tand.
β-Glucosidases	*SUT2-BGL1-HXT2.6*	Chr4: 1774 - 1783	Conv; tand.
β-Glucosidases	*HXT2.5-BGL3-SUT3*	Chr6: 1708 - 1717 *	Tand; conv.
Dityrosine formation	*DIT2-DIT1-DTR1*	Chr2: 1315.9 - 1321.2	Div; Tand
Endo-glucanase	*BGL6-EGC1*	Chr1: 656.5 - 662	Tand.
Endo-glucanase	*HXT2.1-EGC3*	Chr1: 2798.5 - 2803	Div.
Galactose metabolism	*GAL1-GAL10-GAL102-GAL7*	Chr3: 420 - 430	Div; (tand, conv), div.
Histone (H3, H4)	*HHT1-HHF1* *HHF1.1-HHT1.1*	Chr6: 620.9 - 622.8 Chr6: 106 - 106.1	Div.
Histone (H2A, H2B)	*HTB2.2-HTA2* *HTB2.1-HTA1*	Chr6: 643.4 - 645.6 Chr8: 335 - 335	Div.
N-acetyl glucosamine	*NAG4-NAG2-NAG1-NAG5*	Chr6: 11 - 19 *	Conv; div, conv.
Pyrimidine metabolism	*TPN1-THI4-THI13*	Chr3: 1234 - 1238.4	Div; Tand,
	Tandem repeats		
2'-Hydroxyisoflavone reductase	*CIP1.1-CIP1.2-CIP1.3-CIP1.1-CIP1.5*	Chr5: 1530.2 - 1536.5	Tand.
Aldo/Keto reductase	*AKR1-AKR2*	Chr4: 949.8 - 952.4	Tand.
Aldo/Keto reductase	*AKR3-AKR5*	Chr6: 51.7 - 54.2	Tand.
Cinnamyl alcohol dehydrogenase	*CAD3-CAD2*	Chr1: 1857.3 - 1861.2	Tand.
Glutathione S- transferase	*GST1-GST3*	Chr2: 1967.5 - 19700	Tand.
Iron metabolism	*FRE1.1-FRE1.3*	Chr1: 1318.6 - 1324.7	Tand.
Iron metabolism	*FTH1-FRE1.2*	Chr2: 1381.5 - 1386.5	Tand.
L-Rhamnose metabolism	*LRA3-LRA-LRA2-LRA4*	Chr8: 185.8 - 190.5	Tand.
Malate permease	*SSU1-SSU2-SSU3*	Ch3: 1225 - 1230.6	Tand.
Old yellow enzyme (OYE)	*OYE2.5-OYE2.6-OYE2.8*	Chr4: 495.4 - 501.2	Tand, Tand.
Old yellow enzyme (OYE)	*OYE2.9-OYE2.4-OYE2.1*	Chr5: 313.2 - 317.5	Tand.
Peptide transport	*PTR2.1-PTR2.2*	Chr2: 1024 - 1029	Tand.
Taurine catabolism	*IFH2.4-IFH2.3*	Ch1: 661 - 665	Tand.
Urea permease	*DUR3.1-DUR1 (DUR1.2)*	Chr1: 1257.5 - 1276	Tand.

genes share a common 5' sequence, which has some implication with respect to regulation. Divergent genes seem to have evolved in longest association with one another, since transcription might proceed in either direction from a central shared promoter sequence (Jeffries and Van Vleet 2009).

Some of the gene clusters have a tandem repeat conformation. For example, the OYE gene family, which codifies for NADPH dehydrogenase, is present in 12 members, six of which appear in two triplet clusters of tandem repeats, and phylogenetic analysis suggests that a triplicate cluster underwent three rounds of duplications, followed by differentiation with little or no loss of intermediate activities. Similarly, the genes in the AKR family of aldo/ketoreductases is present in six members, with two paired clusters; cinnamyl alcohol reductases (CAD) are present in five members with two of them (*CAD2* and *CAD3*) in a tandem, and 2-hydroxyisoflavone reductases also present five members (*CIP1.1-CIP1.5*). The presence of *CAD* and *CIP* genes is not common in related genomes, and their function is poorly characterised, so these activities seem to be specific to *S. stipitis* (Jeffries and Van Vleet 2009).

1.5.3 *S. stipitis* for ethanol fermentation

There are three stages in the conversion of xylose to ethanol: (i) Xylose transport and subsequent reactions to enter the pentose phosphate pathway (PPP) (ii) Non-oxidative reaction of the PPP (iii) Glycolysis. Despite the pathways for pentose fermentation has been already studied and established, the specific characteristics of *S. stipitis* are not completely known.

The rate limiting step in ethanol fermentation in *S. stipitis* is the transport of sugars into the cells (Ligthelm *et al.* 1988). When the concentration of sugars is high, low affinity transport systems are used, whereas when the concentration of sugar is low, the high affinity systems are preferred. Low affinity systems share the transport between glucose and xylose, and the transport of xylose is inhibited by glucose by non-competitive inhibition (Kilian and van Uden 1988). Moreover, the rate of glucose consumption is higher than the xylose consumption under similar growth conditions (Agbogbo *et al.* 2006). Both implications seem to indicate that glucose is the preferred carbon source.

Three genes have been identified as important sugar transporters in *S. stipitis*: *SUT1, SUT2* and *SUT3. SUT2* and *SUT3* are very similar to the glucose transporters found in *S. cerevisiae*. They exhibit more affinity for glucose and are expressed only under anaerobic conditions but independently of the carbon source. On the other hand, *SUT1* is expressed independently of the oxygen supply, and its disruption eliminates the low affinity transport of xylose (Weierstall, Hollenberg and Boles 1999). Moreover, a

naturally-occurring specific xylose transporters with no glucose uptake activity (Xyp29) has also been identified (Du, Li and Zhao 2010).

Once introduced in the cell, the first step for xylose metabolism is its reduction to xylitol by xylose reductase (XR, *XYL1*), which shows affinity for both NADH and NADPH, although its higher towards NADPH (Verduyn *et al.* 1985). Then xylitol is converted to xylulose by xylitol dehydrogenase (XDH, *XYL2*), which shows affinity only for NAD^+ (Metzger and Hollenberg 1995). Finally, a xylulokinase (XK, *XYL3*) converts xylulose to xylulose-5-phosphate (Jin *et al.* 2002), which is an intermediate of the PPP. The expression of XR and XDH is induced in the presence of xylose, but their activity is repressed in the presence of glucose (Bicho *et al.* 1988). Hence, glucose concentration acts as a rate limiting step in xylose fermentation. The expression of XK is induced by xylose, and once it is expressed it does not seem to be rate limiting (Jin *et al.* 2002).

These first reactions normally cause a cofactor unbalance, since the NADPH required by the xylose reductase is not re-generated, which leads to xylitol accumulation in most natural xylose-fermenting yeasts (Jeffries 2006). However, xylitol generation in *S. stipitis* is minimum for two main reasons. First, *S. stipitis* is one of the few types of yeasts in which XR can use both NADPH and NADH as cofactors, which means that the activity of the enzyme is not compromised when NADPH levels are low (Agbogbo and Coward-Kelly 2008) (**Figure 1.14**).

Second, *S. stipitis* presents an efficient system for the regeneration of NADPH (Balagurunathan *et al.* 2012), which involves the catalysis of 2-oxoglutarate to succinate. First, 2-oxoglutarate is converted to L-glutamate by glutamate dehydrogenase (*GDH2*) consuming NADH, then, the L-glutamate is decarboxylated to 4-aminobutyrate by a glutamate decarboxylase (*GAD2*) and this is subsequently transformed to succinate semialdehyde by an aminotransferase (*UGA1)*, and this is finally oxidised to succinate by succinate semialdehyde dehydrogenase (*UGA2*) using $NADP^+$. Therefore, the net result of the reactions is consumption of NADH and production of NADPH (Jeffries *et al.* 2007; Jeffries and Van Vleet 2009) (**Figure 1.14**).

Once the different sugars have been introduced in the cell, glucose will be degraded to glyceraldehyde-3-phosphate through the glycolysis. On the other hand, xylose will be transformed to xylulose-5-phosphate, and this will be either degraded to glyceraldehyde-3-phosphate via the PPP, or transformed into fructose-6-phosphate and then degraded through glycolysis. The glyceraldehyde-3-phosphate produced will be subsequently catabolised to pyruvate, and this one transformed to acetaldehyde by the pyruvate decarboxylase. Finally, acetaldehyde will be used to produce ethanol by alcohol

dehydrogenases (**Figure 1.14**). These pathways are conserved from *S. cerevisiae* to *S. stipitis* (Kwak and Jin 2017).

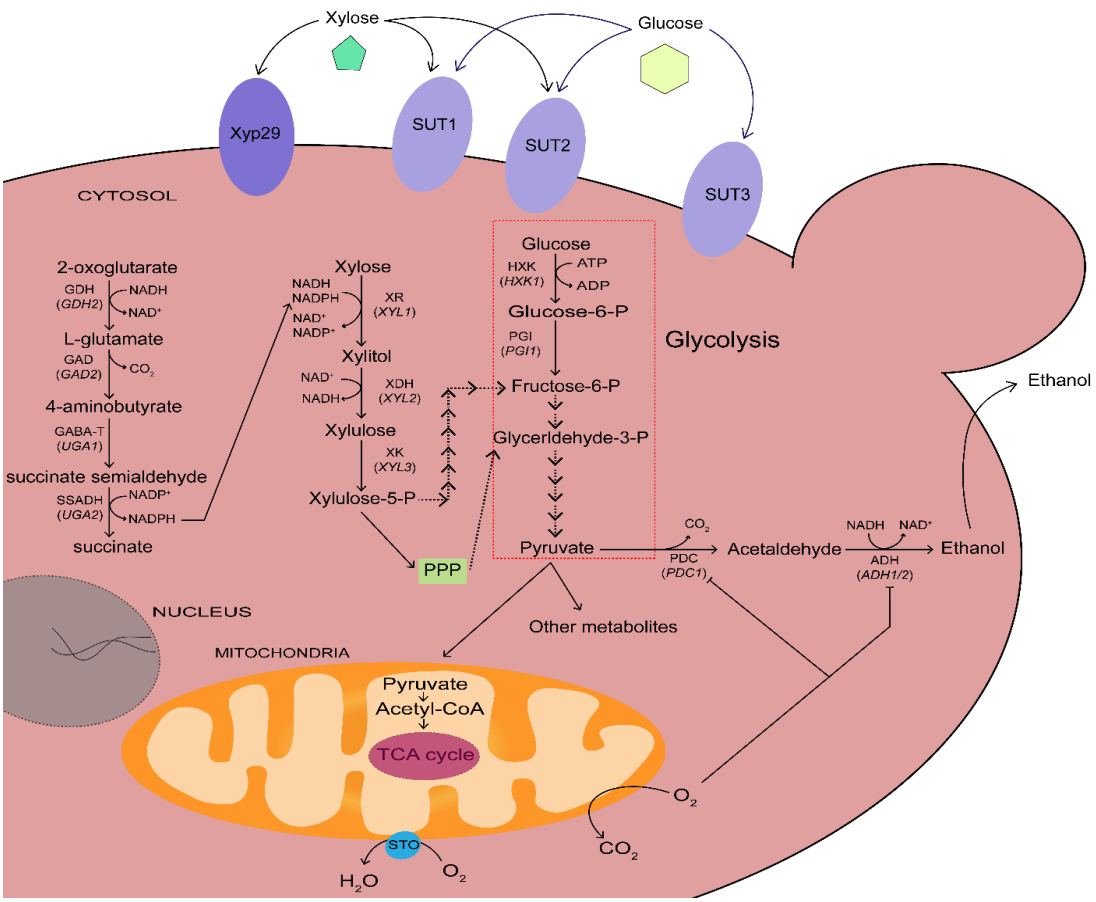


Figure 1.14. Metabolic pathways involved in ethanol fermentation from glucose and xylose in *S. stipitis*. Abbreviations: ADH: Alcohol dehydrogenase, GABA-T: 4-aminobutyrate transaminase, GAD: Glutamate decarboxylase, GDH: Glutamate dehydrogenase, HXK: Hexokinase, PDC: Pyruvate decarboxylase, PGI: Phosphate isomerase, PPP: Pentose phosphate pathway, SSADH: Succinate semialdehyde dehydrogenase, STO: Sham-sensitive terminal oxidase, SUT: Sugar transporter, TCA: Tricarboxylic acid cycle (Krebs cycle), XDH: Xylitol dehydrogenase, XK: Xylulokinase, XR: Xylose reductase, Xyp: Xylose transporter. Red dashed lines: Glycolysis reactions. Black dashed lines: Not shown metabolic reactions.

Nevertheless, *S. stipitis* is a Crabtree negative yeast (Prior BA, Kilian SG and Du Preez JC 1989), which means that is able to ferment sugars only when the oxygen content is limited, independently of the sugar concentration (unlike *S. cerevisiae*, which, as a Crabtree positive yeast, is able to ferment sugars when its concentration is very high, even in the presence of oxygen) (Papini *et al.* 2012). This difference in behaviours is related to the induction of enzymes responsible of ethanol production (alcohol dehydrogenases (ADH, *ADH1* and *ADH2*) (Cho and Jeffries 1998; Cho and Jeffries 1999; Klinner *et al.* 2005), aldehyde dehydrogenase (AlDH) and the pyruvate decarboxylase complex (PDC) (Lu, Davis and Jeffries 1998)) when the oxygen tension is reduced.

One of side effects of low oxygen levels during fermentation for Crabtree negative yeasts is the suppression of respiration and the generation of ATP by the electron transport chain (Cho and Jeffries 1999) in the cytochrome c respiration pathway.

These limitations are partially overcome in *S. stipitis* due to the presence of an alternative respiration pathway, known as the SHAM-sensitive respiration system (Jeppsson, Alexander and Hahn-Hagerdal 1995), or STO pathway, in addition to the standard electron transport chain in which acts the cytochrome c oxidase (Cox). This pathway, based on the presence of an alternative terminal oxidase (Sto) that is sensitive to salicylhydroxamic (SHAM) acid, has also been identified in other yeasts (Aoki and Ito-Kuwa 1984; Viola *et al.* 1986; Poinsot *et al.* 1987), fungi (Lambowitz and Slayman 1971) and plants (Douce and Neuburger 1989).

The STO pathway allows the cell to maintain electron transfer, since Sto can accept electrons from both independent NADH and succinate dehydrogenases, although it does not seem to be related with ATP production in *S. stipitis,* since these dehydrogenases are not involved in proton translocation (Shi *et al.* 2002). Therefore, the pathway is also involved in the resolution of the cofactor unbalance commonly observed in transformation of xylose to xylulose-5-phosphate (Jeppsson, Alexander and Hahn-Hagerdal 1995) (**Figure 1.14**). Moreover, Sto has been reported to show more affinity for oxygen (Lambers 1982), so it might act as a scavenging agent for residual oxygen when its concentration is low.

Despite the metabolic advantages described, the production of ethanol by natural isolates of *S. stipitis* have reached levels of 61 g/L using xylose as carbon source (Slininger *et al.* 2006) and 0.406 $g_{ethanol}$ / g_{xylose} (88% of the theoretical maximum production (0.46 $g_{ethanol}$ / g_{xylose}) (Lee, Tae-Hee *et al.* 2001)) in an optimised medium, or 42 g/L (85% of ethanol maximum accumulation) for strains adapted to high solids hydrolysates (Slininger *et al.* 2015), these are still far from the levels obtained by *S. cerevisie* from glucose (around 80 g/L (Kasavi *et al.* 2012) with a theoretical maximum yield of 0.511 $g_{ethnaol}$ / $g_{glucose}$) (Lee, Tae-Hee *et al.* 2001). Hence, increasing the rate of fermentation of xylose by *S. stipitis* could improve its use for commercial applications (Jeffries and Van Vleet 2009).

Although the extensive knowledge in ethanol fermentation by *S. stipitis* is poor compared to *S. cerevisiae,* plenty of studies have been conducted to optimise growth and fermentation conditions. The optimal temperature for fermentation is between 25 and 35 ºC, and the optimal pH is in the range of 4.0-7.0 (du Preez, Bosch and Prior 1986; Slininger *et al.* 1990), although these values seem to be strain dependent. The nutrients on the media also play an important role on ethanol production. Addition of amino acids

and nitrogen sources increases the non-growth associated ethanol production (Slininger *et al.* 2006), and ammonium salts increase the ethanol productivity and ethanol to biomass yield (Guebel *et al.* 1992; Agbogbo and Wenger 2006). Low levels of magnesium result in xylitol accumulation and high NADH content (Mahler and Nudel 2000), although magnesium ions and amino acid enrichment of urea as a nitrogen source enhance both ethanol and furan tolerance (Slininger, Gorsich and Liu 2009).

Despite of showing high ethanol productivity and low levels of xylitol as a by-product, genetic engineering techniques have been applied to obtain strains with improved fermentation traits or for the accumulation of other by-products of interest. Nevertheless, the efficiency of genetic manipulation techniques relies strongly on the dominating pathway for DSB DNA repair. A vast majority of techniques rely on the use of DNA templates that will be used to repair the break by homologous recombination. Transformation of *S. stipitis* is relatively complicated, especially when compared to *S. cerevisiae,* mainly because the main pathway to repair DSBs is NHEJ (Maassen *et al.* 2008). This leads to an increase in the random chromosomal integration of the DNA template, leading to mutagenesis. Moreover, *S. stipitis* has been reported to be resistant to most commonly used antibiotics (*Biotechnology of Yeasts and Filamentous Fungi* 2017), and the possible markers to be used require codon adaptation to account for the CTG clade phylogeny. Despite of this, several genetic modification methods have been developed and adapted: sexual mating and sporulation (Melake, Passoth and Klinner 1996), electroporation (Yang *et al.* 1994), lithium acetate (Cho and Jeffries 1999), freeze transformation (Passoth, Hahn-Hägerdal and Klinner 2003), loxP/Cre recombination system (Laplaza *et al.* 2006), and more recently CRISPR-meditated genome edition (Cao *et al.* 2018) have been successfully applied. This has allowed obtaining a number of auxotrophic mutant strains by combination of chemical mutagenesis and genetic modification (*ura3* (Yang *et al.* 1994)*, leu2* (Lu *et al.* 1998, p.2)*,* or *trp5* and *his3* (Jutta Hagedorn 1990)), although the availability of these auxotrophic mutants is still limited.

Genetic engineering has been successfully applied to generate mutants able to grow and ferment xylose in anaerobic conditions (Shi and Jeffries 1998). Strains with improvements in biomass and ethanol production have also been obtained (Den Haan and Van Zyl 2003; S.-H. Chen *et al.* 2012). Moreover, strains defective in *XYL2* have been obtained for high production levels of xylitol (Ko, Kim and Kim 2006; R. C. L. B. Rodrigues *et al.* 2007), and strains with the ability of efficient production of lactic acid from xylose have been constructed through the expression of the lactate dehydrogenase gene (*LDH*) of *Lactobacillus helveticus* under the control of the yeast promoter *ADH1* (Ilmén *et al.* 2007). Moreover, recombinant strains that produce fumaric acid from xylose

(Wei *et al.* 2015), shikimate (Gao *et al.* 2017), or that accumulate S-adenosylmethionine have also been generated (Križanović and Butorac 2015).

Despite of the improvements in fermentation rates achieved by genetic engineering, there are still some limitations in *S. stipitis* that complicate its use for industrial scale up: (i) Low rate of fermentation when compared to *S. cerevisiae* (Jeffries *et al.* 2007) (ii) Higher susceptibility to ethanol inhibition when compared to *S. cerevisiae* (du Preez, van Driessel and Prior 1989) (iii) Oxygen is required for growth but it has to be limited for fermentation (Papini *et al.* 2012) (iv) Poor tolerance to inhibitors present in lignocellulose hydrolysates (Bellido *et al.* 2011) (v) Lack of simultaneous fermentation of glucose and xylose, being glucose the preferred carbon source (Agbogbo *et al.* 2006) (vi) Ethanol reassimilation, which implies that the ethanol produced will be used as carbon source, instead of the xylose remaining in the medium (Skoog *et al.* 1992).

Several methodologies have been applied to overcome the limitations of the use of *S. stipitis* for ethanol fermentation. One of the most popular is the adaptation of strains by repeated sub-culturing, or by biomass recycling while increasing the concentration of selective pressures (for example: inhibitors, temperature, or sugar concentration) (Amartey and Jeffries 1996; Huang *et al.* 2009; Yang *et al.* 2011). ALE experiments can be substituted by other procedures, such as random mutagenesis by UV exposition (Bajwa *et al.* 2009), genome shuffling via protoplast fusion (Bajwa, Pinel, Vincent J.J. Martin, *et al.* 2010; Bajwa *et al.* 2011), protoplast fusion with other yeasts, such as *S. cerevisiae* (Jetti *et al.* 2019), or even genome shuffling by transformation of a yeast with the total DNA isolated from another yeast specie (Zhang and Geng 2012). Moreover, combinations of both adaptation and subsequent mutagenesis are also commonly applied (Watanabe *et al.* 2011; Harner *et al.* 2015; Slininger *et al.* 2015).

However, the only reported re-sequencing study of an improved strain of *S. stipitis* detects non-synonymous SNPs in only 12 ORFs (of which 1 was in a gene related to ethanol fermentation, *ALD7),* but their relation with better ethanol production was not further studied (D. R. Smith *et al.* 2008). Despite of this, differences in the genome structure of natural isolates of *S. stipitis* have been reported by pulse field gel electrophoresis (PFGE) studies (Passoth *et al.* 1992), which indicates relative genome plasticity. These observations have also been reported and deeply studied in other CTG clade yeasts, such as *C. albicans* (Rustchenko-Bulgac 1991)*, C. glabrata* (Poláková *et al.* 2009) or *D. hansenii* (Petersen and Jespersen 2004). Moreover, changes in the ploidy level as a response to stress have also been reported for *S. stipitis* (Talbot and Wayman 1989), which is an indication of genome modification events as a result of stress

adaptation processes. Nevertheless, the mechanisms involved in these genome variations and their effects are not known.

1.6 AIMS

The adaptation of fermentation procedures and genetic engineering protocols have allowed the isolation of *S. stipitis* strains with improved lignocellulose conversion phenotyoes, and highlighted its potential use for industrial applications. However, the lack of knowledge on the genome modifications of these strains complicates the determination of the characteristics required in a strain for efficient fermentation profiles. Hence, the study of genome changes in adapted strains of *S. stipitis* might help to determine which modifications are fundamental to obtain better fermenting strains.

The hypothesis of this study is that natural isolates of *S. stipitis,* as other yeasts belonging to the CTG clade, are characterised by extensive genome diversity and that this variability might drive the development of industrial-relevant phenotypes that could be exploited for more efficient bioethanol production. This will be addressed in several steps:

1. To demonstrate whether the genome of *S. stipitis* is plastic, the genomic conformation of a library of natural isolates (27) obtained from different habitats and countries will be analysed by karyotyping.
2. The genome of the strain Y-11545, sequenced by Jeffries *et al*. (2007), and publicly available in NCBI, will be analysed to find repetitive regions, since they are determinant in the plasticity observed in other yeasts of the CTG clade.
3. The genome of a natural isolate strain that exhibits karyotypic differences with Y-11545 will be sequenced and compared to assess whether repetitive regions have conditioned these genomic variations during strain evolution.
4. The effect of genome plasticity on different traits desirable for industrial fermentation will be assessed.
5. The connection between phenotypic and genomic differences will be evaluated in strains diverged from a short-term evolution experiment. Strains with improved fermentation phenotypes will be studied by karyotyping and their differences will be determined by whole genome sequencing. This will also allow to assess if short-term evolution is also conditioned by repetitive regions in the genome of *S. stipitis.*

Chapter 2

Materials and methods

2.1 Strains used

The strains used for this project were obtained from two collections: Agricultural Research Service (ARS) Collection, run by the Northern Regional Research Laboratory (NRRL) (Peoria, Illinois, USA), and National Collection of Yeast Cultures (NCYC) (Norwich, United Kingdom). The freeze and dried biomass was re-suspended in YPD and grown overnight at 30 ºC with agitation at 200 rpm. The overnight was used to make 40% glycerol stocks and stored at -80 ºC. A list of all the strains used is found in **Table 2.1**.

2.2 Growth medias

Strains were grown in different medias depending on the experiment to be conducted. Routine culture was performed on Yeast Extract-Peptone-D-Glucose (YPD) media, whereas growth evaluation was conducted on SC using glucose (SC-Glucose), xylose (SC-Xylose), or a mixture of 60% glucose and 40% xylose (SC-Mix). Uridine and adenine hemisulfate were added as growth supplements. Solid media was prepared by addition of agar (**Table 2.2**). Once prepared, the media was autoclaved.

Table 2.2. Composition of the growth medias used in this project.

Compound	YPD	SC-Glucose	SC-Xylose	SC-Mixture
Glucose (g/L) (Fisher, #G/0500/61)	20	20	-	12
Xylose (g/L) (Merk, #W360600)	-	-	20	8
Bactopeptone (g/L) (BD Biosciences, #211677)	20	-	-	-
Yeast Extract (g/L) (Oxoid, #LP0021)	10	-	-	-
Kaiser Complete SC Mixture (g/L) (Formedium®, DSCK1000)	-	2.002	2.002	2.002
Yeast Nitrogen base without aminoacids (g/L) (Difco®, #291940)	-	6.7	6.7	6.7
Uridine (g/L) (Merk, #U3750)	0.08	0.08	0.08	0.08
Adenine hemisulfate (g/L) (Merk, #A3159)	0.05	-	-	-
Agar (g/L) (Melford Biolaboratories, #9002-18-0)	20	20	20	20

To assess the effects of growth inhibitors commonly present in lignocellulose hydrolysates SC-Mix was used with the additions of different compounds in two conformations possible: mild inhibitory conditions and strong inhibitory conditions (**Table 2.3**). The media was prepared as SC Mix, then all the compounds were added, and pH adjusted at 6.8±0.1 with 10M NaOH and then filter sterilised (Thermo Scientific™ Nalgene™ Rapid-Flow™ sterile disposable bottle top filters with SFCA Membrane, #10139560, 0.2 µm pore size).

Table 2.3. Compounds used for the preparation of the inhibitor cocktail used in this study.

Inhibitor	Mild inhibitory conditions	Strong inhibitory conditions
Acetic acid (g/L)	0.9	6.1
Formic acid (g/L)	0.09	12.4
Furfural (g/L)	0.1	1.32
4-hydroxybenzoic acid (g/L)	0.005	0.007
Catechol (g/L)	0.009	0.042
Vanillin (g/L)	0.063	0.086
Vanillic acid (g/L)	0.011	0.084

2.3 Colony PCR and gel electrophoresis

Strains were streaked in YPD agar plates and grown overnight at 30 ºC. A single colony was peaked with a toothpick and re-suspended in 40 µl of 0.02M NaOH (Fisher Scientific™, #S/4920/60). Then, the suspension was incubated at 100 ºC for 10 minutes and put on ice for other 10 minutes. Finally, the cells were centrifuged at 2000 rpm for 5 minutes, and the supernatant was used for PCR as DNA extract. The PCR reaction was conducted in a C1000™ Thermal Cycler (Biorad®), following the conditions stated in **Tables 2.4 and 2.5**. PCR amplicons were analysed in an 1% agarose gel (Melford Biolaboratories, #MB1200) in 0.5x TBE, with ethidium bromide (0.5 µg/ml) and visualised under UV light using a Syngene GBox Chemi XX6 Gel imaging system.

Table 2.4. Reagents used for PCR reactions and their final concentration.

	Final concentration	Volume for 1x Master Mix reaction
Buffer (PCR Biosystems® PB10.15-06)	1x	1.50 µl
dNTPs (10 mM stock of mixture of dATP, dCTP, dGTP, dTTP, (VWR))	333 µM	0.50 µl
Primer forward (10 µM stock)	400 nM	0.60 µl
Primer reverse (10 µM stock)	400 nM	0.60 µl
DNA Taq Polymerase (PCR Biosystems® PB10.15-06) (6000 U)	52 U	0.13 µl
DNA (sodium hydroxide extraction)	-	1 µl
MQ water	-	Up to 15 µl

Table 2.5. PCR protocol.

	Temperature (ºC)	Time (s)
Initial denaturation	94	300
Denaturation	94	45
Annealing	Primer depending	15
Extending	72	60 (per Kylobase)
Final extension	72	300

2.4 Determination of *S. stipitis* species

To confirm genetically that all the strains are *S. stipitis*, the sequence within the ribosomal DNA region spanning the internal transcribed spacers ITS1, 5.8S and ITS2, and the sequence of the D1/D2 domain of the 26S rDNA gene were used. The DNA was extracted following the protocol explained in 2.3, and the region was then amplified using the primers AB796 and AB797, and AB798 and AB799 respectively (**Table 2.6**) as previously proposed by Villa-Carvajal *et al*. (Villa-Carvajal, Querol and Belloch 2006). The PCR product was analysed by PCR electrophoresis, purified using the E.Z.N.A Cycle Pure Kit (V-Spin Column) (VWR, #D6492-02), and sent to sequence to Eurofins Scientific using primers AB796 and AB798 for each region studied.

2.5 CHEF electrophoresis and Southern Blotting

2.5.1 Solutions required

100 mM and 50 mM EDTA were prepared fresh from an EDTA 0.5M stock. Zymolyase® 100T from Arthrobacter luteus (Amsbio, # 120493-1), was prepared as a stock of 10 mg/ml in 50% glycerol. LET was prepared by mixing 200 ml of 0.5M EDTA, 10 ml of 1M Tris pH=8.0 and adjusting to 1L with water. NDS was prepared by mixing 200 ml of 0.5M EDTA, 50 ml of 1M Tris pH=8.0, 50 ml 10% SDS and adjusting to 1L with water. The mixture NDS-Proteinase K must be prepared before use (1 mg of enzyme per 5 ml of NDS).

For Southern blotting, several reagents are required. 20x SSC is prepared mixing 75.3 g of NaCl and 88.2 g of trisodium citrate, for a final volume of 1L. The final pH = 7.0 was adjusted with HCl. Stock concentrations of 10% SDS, 4M NaOH and 5M NaCl were also needed. Maleic acid buffer was prepared as a mixture of Maleic acid and NaCl to a final concentration of 100 mM and 150 mM, respectively. NaOH pellets were used to adjust to pH = 7.5. Blocking buffer (DIG Easy-Hyb blocking reagen, #11096176001) was prepared as 10x in maleic acid buffer according to manufacturer instructions. Buffer 3 was prepared by mixing 300 ml of maleic acid buffer and 900 µl of Tween 20 until this was totally solved. Detection

buffer consisted in a mixture of Tris base and NaCl to a final concentration of 100 mM each and adjusted to pH = 9.5 using HCl.

2.5.2 Plug preparation

Intact yeast chromosomal DNA was prepared following the method proposed by Schwartz and Cantor (Schwartz and Cantor 1984) with modifications. Freshly streaked cells of *S. stipitis* were grown overnight in 3 ml of YPD at 30 ºC, and the volume of cells corresponding to an OD_{600} of 7 was centrifuged at 3000 rpm for 3 minutes at room temperature. The cells were then washed with 1 ml of 50 mM EDTA pH=8.0 and centrifuged in the same conditions. After complete removal of the EDTA, 20 µl of 10 mg/ml Zymolyase was added, and immediately after, 300 µl of 1% low melt agarose (Biorad®, #1613112) in 100 mM EDTA was used to carefully re-suspend the cells. The mixture was then transferred to the plug mould. Once the plugs solidified (from 15 to 45 minutes) they were transferred to tubes with 2.5 ml of LET, and subsequently, 25 µl of ß-mercaptoethanol (Merk, #M6250) were added. The plugs were incubated overnight at 37 ºC in a water bath.

After that, the LET solution was removed and the plugs were washed twice (30 minutes each) with 3 ml of 50 mM EDTA pH=8.0. Then, 3 ml of NDS-Proteinase K solution was added to the plugs, which were subsequently incubated at 50 ºC in a Roller-Blot Hybridizer HB-30 (Techne®) for two nights.

Finally, the NDS solution was removed and the plugs were washed with 3 mL of 50 mM EDTA pH=8.0 for 30 minutes and resuspended in fresh 50 mM EDTA pH=8.0. The plugs were then immediately used for electrophoresis or stored at 4 ºC until used.

2.5.3 Genomic DNA separation by CHEF electrophoresis

The plugs were run in CHEF DR II system (Biorad®) in a 1% megabase agarose gel (Biorad®, #1613110) in 2L of 0.5x TBE at 14 ºC with the following settings: 60-120s switching time for 12 hours at 6 V/cm, followed by a switching time of 120-300s for 26 hours at 4.5 V/cm. The gel was stained in 0.5x TBE with ethidium bromide (0.5 µg/ml) for 30 minutes and washed in water for other 30 minutes. The gel was then visualised under UV light using a Syngene GBox Chemi XX6 Gel imaging system.

2.5.4 Southern Blotting

After running, staining, and imaging the gel, several washes were conducted for Southern Blotting. First, the gel was rinsed in 250 ml of 0.25M HCl for 15 minutes, and then in water for 5 minutes. After that, the DNA was denatured by gently washing the gel in 1.5M NaCl + 0.5M NaOH for 15 minutes, twice. Subsequently, the gel was rinsed in water for 5 minutes,

neutralised by gently rocking in 1M Ammonium acetate + 0.02M NaOH for 30 minutes, twice, and rinsed in 20x SSC for 10 minutes. Finally, the DNA was transferred from the CHEF megabase agarose gel to a Zeta-Probe GT Membrane (Biorad®, #162-0196) in 20x SSC overnight, and crosslinked using UV light (150 mJ) with the CROSSLINKER® CL-508 (UVITEC Cambridge). The membranes were then immediately used or stored in dark and dry environments.

Prior to use, the membranes were incubated with 50 ml of DIG Easy Hyb (Roche®, 11603558001) for 1 hour at 42 ºC. Then the probe designed to target the chromosome of interest was boiled at 98 ºC for 10 minutes, and 100 µl was added to 50 ml of DIG East Hyb. The membrane was then incubated in this mixture overnight at 42ºC with gentle agitation. The probe, designed to target the chromosome 5 of *S. stipitis* Y-11545, was amplified by PCR from genomic DNA with the primers AB1028 and AB1029 (**Table 2.6**) using DIG-dUTP (Roche®, #11573152910) as label.

After the incubation, the membrane was rinsed with vigorous shaking in 2x SSC + 0.1% SDS for 10 minutes, twice. Then, it was washed in pre-warmed 0.5X SSC + 0.1% SDS for 20 minutes, twice, and after that rinsed in maleic acid for 5 minutes. The membrane was subsequently blocked using 1x Blocking buffer for 3 hours, and immediately after, it was incubated for further 4.5 hours with a mixture of 5 µl of anti-digoxigenin-Alkaline Phosphatase antibody (Roche®, #11093274910) with 50 ml of 1x Blocking buffer. After the incubation, the membrane was washed in 150 ml of buffer 3 for 15 minutes, twice, and then rinsed in detection buffer for 5 minutes first, and for further 2 minutes after. Finally, CDP Star ready to use (Roche®, #12041677001) was used for DNA detection according to manufacturer instructions and the membrane was developed in Optimax 2010 X-ray film processor (Euroteck Systems UK Ltd).

2.6 Genome analysis

2.6.1 DNA-sequencing library preparation

DNA was extracted from an overnight culture in YPD using the QIAGEN genomic tip 100/G kit (Qiagen®, #10243) according to manufacturing protocol. DNA concentration was evaluated using a NanoDrop™ 1000 Spectophotometer (Thermo Scientific) and fragment size was evaluated using the TapeStation 4200 (Agilent technologies)

For long-read sequencing, MinION (Oxford Nanopore, Oxford UK) was used. The DNA library was prepared using the Ligation Sequencing Kit 1D (SQK-LSK108) according to manufacturing protocol and run on the Oxford Nanopore MinION flowcell FLOMIN 106D.

For Illumina libraries the DNA was sheared using the Covaris M220 with microTUBE-50 (Covaris 520166) and size selected using the Blue Pippin (Sage Science). The library was constructed using the PCR-free method using NEBNext End Repair (E6050S), NEBNext dA-tailing (E6053S) and Blunt T/A ligase (M0367S) New England Biolabs modules. Libraries were sequenced using Illumina Miseq V3 2x300bp PE (MS-102-3003).

Both sequencing libraries were prepared using the same DNA extraction.

2.6.2 Genome assembly

A summary of the pipeline used for genome assembly is presented in **Figure 2.1**. The codes used for this project can be found in Github (https://github.com/harrisonlab/pichia).

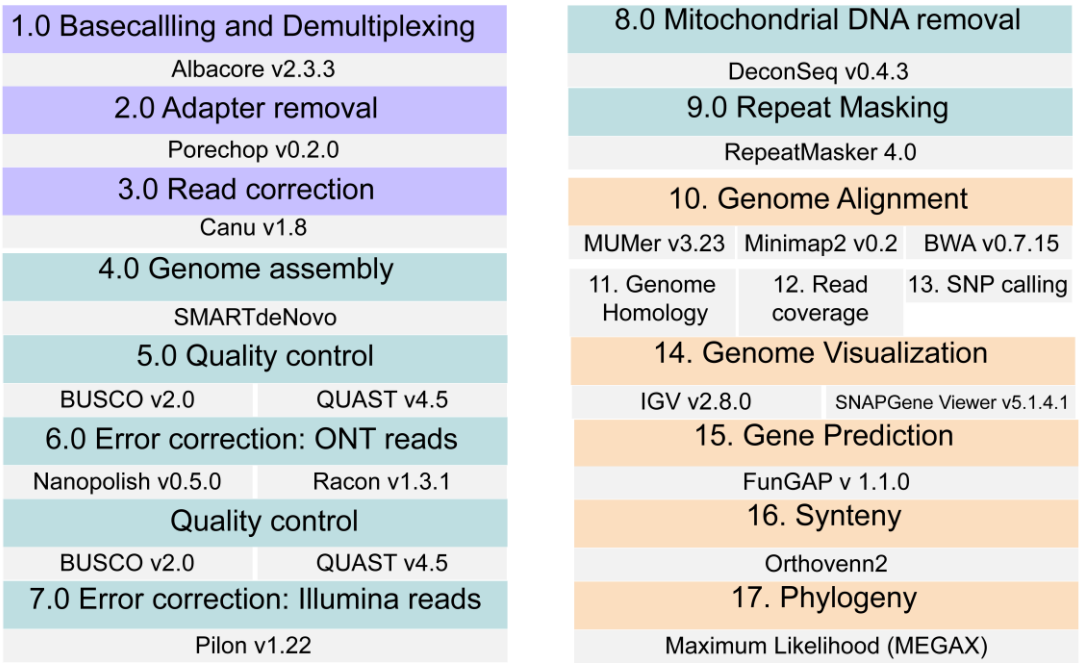


Figure 2.1. Pipeline used for the assembly of the genomes sequenced during this project. Purple colour indicates the data filtering and pre-processing steps. Blue indicates the *de novo* genome assembly steps. Orange indicates assembly analysis.

Prior to generate the genome assembly, the sequencing reads were processed. Basecalling and demultiplexing was conducted using albacore v2.3.3 (Wick, Judd and Holt 2019), and the adapters were removed with Porechop v0.2.0 (https://github.com/rrwick/Porechop#how-it-works) (Wick *et al.* 2017). Finally, Canu v1.8 was used for the error correction and trimming of the ONT reads (Koren *et al.* 2017).

The genome assembly was generated using the SMARTdenovo assembler (https://github.com/ruanjue/smartdenovo). Subsequently, ONT raw reads were used for both error correction with Racon v1.3.1 (https://github.com/isovic/racon) (Vaser *et al.* 2017) and for

polishing with Nanopolish v0.5.0 (https://github.com/jts/nanopolish/blob/master/README.md). Finally, the assembly was polished using Illumina sequencing reads with Pilon v1.22 (Walker *et al.* 2014). The quality of the assembly was calculated after each one of the correction and polishing steps with BUSCO v2.0 (Benchmarking Universal Single Copy Orthologs) using the database of Saccharomycetales_odb9 (Waterhouse *et al.* 2013; Simão *et al.* 2015), and with Quast v4.5 (Gurevich *et al.* 2013).

Once the assembly was finished, the mitochondrial DNA was removed with DeconSeq v0.4.3 (Schmieder and Edwards 2011), using the *C. albicans* mitochondrial DNA as a template. Finally, the repeats were masked with RepeatMasker 4.0 (Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0., http://www.repeatmasker.org).

2.6.3 Genome analysis

The final assembly after the repeat masking step (in contigs) was analysed by BLASTN to reveal the presence or absence of centromeres, according to the sequences proposed by Coughlan and Wolfe (Coughlan and Wolfe 2019). Further improvements in the genome assembly to obtain a chromosome level assembly. BLASTN was used to discover overlapping regions between the different contigs and determine which ones could be assembled as a unique chromosome. The assembly was finished to a chromosome level manually using the SNAPGene Viewer v5.1.4.1 software (from GSL biotech; available at snapgene.com). SNAPGene Viewer was also used for genome visualization.

The completely assembled genomes (Y-7124, Y-50859 and Y-50861) were aligned against each other and against the reference strain Y-11545 using the program Nucmer, inside Mummer v3.23 (Kurtz *et al.* 2004) to unveil the presence of genome modifications.

Minimap2 v0.2 (Li 2018) was used for the alignment of the raw ONT reads against the genomes assembled and the reference strain for subsequent study of the read coverage with SAMtools v1.3 (http://samtools.sourceforge.net) (Li *et al.* 2009). Both type of alignments were plotted using Circos v0.69.4 (Krzywinski *et al.* 2009) and Rstudio (R version 4.0.2 (http://www.R-project.org/)). SAMtools was also used for sorting and indexing of the ONT reads alignment over the genomes assembled, and bedtools v2.26.0 for the alignment transformation from bam to fasta format (Quinlan and Hall 2010). The alignments in fasta were subsequently visualised with the software Integrative genome viewer v2.8.0 (IGV, from the Broad Institute) (Robinson *et al.* 2017).

The alignment of all the Illumina reads from all the sequenced isolates and the reference strain was conducted with BWA-MEM (Burrows-Wheeler Alignment) version 0.7.15 (Li 2013). This was used for both the determination of the Illumina read coverage over the assembly and for the determination of variants (SNPs).

The variant calling was conducted with Genome Analysis Toolkit (GATK4) (McKenna *et al.* 2010) (Broad Institute, https://gatk.broadinstitute.org/hc/en-us/articles/360036194592-Getting-started-with-GATK4).

Synteny analysis were conducted using Orthovenn2 (Xu *et al.* 2019). The gene models of the strains sequenced in this study were generated with FunGAP v 1.1.0 (Min, Grigoriev and Choi 2017), whereas for the reference strain Y-11545 the models published by Jeffries *et al.* (Jeffries *et al.* 2007) were used.

Finally, the phylogenetic analysis of the ITS5.8S region was conducted using the nucleotide sequences determined during this project (See section 2.4) and from several organisms available at NCBI. The study of the phylogenetic divergence in transposons was analysed by comparison of the amino acid sequence of retrotranscriptase domain in the POL ORF present in non-LTR transposable elements of several organisms. All the sequences analysed in this study were aligned by Clustal-Wallis using either Jalview v2.11.1.0 (Waterhouse *et al.* 2009) or MEGAX (Kumar *et al.* 2018). The phylogenetic trees were constructed using the maximum likelihood algorithm with 100 bootstraps, using MEGAX (Kumar *et al.* 2018).

The domains and motifs of different proteins were determined *in silico* using SMART (The European Molecular Biology Laboratory)(Letunic and Bork 2018)

2.6.4 Validation of the genome modifications

The translocation between chromosome 5 and 7 in the strains Y-11545 and Y-7124 were assessed via Southern Blotting, using a probe that hybridised the chromosome 5 of both strains, generated by amplification of genomic DNA with the primers AB1028 and AB1029 (**Table 2.6**). The translocation between chromosome 1 and 2 between the strains Y-7124 and Y-50859 was assessed via PCR using the primers AB966, AB967, AB968 and AB969 (**Table 2.6**). The aneuploidy was demonstrated via: (i) CHEF electrophoresis (See section 2.5) with the following settings: 60-12s switch time for 12 hours at 6V/cm followed by a 120-300s switch time at 4.5V/cm for 12 hours (ii) Southern blotting using the probe designed for chromosome 5 (See section 2.5), and (iii) PCR expanding the area in which telomeric repeats and high coverage region in chromosome 5 unite, using the primers AB1015 and AB1019 (**Table 2.6**).

2.7 Fluctuation analysis

The rate of isochromosome 5L (i-5L) loss was measured by fluctuation analysis. The strain Y-50859 was grown overnight in 5 ml of YPD, plated at a cell density of 100 cells and grown 48 hours at 30 ºC. Then 7 colonies in which the presence of the of the aneuploidy was confirmed via PCR (See section 2.6.4) were grown for 15 hours in 5 ml of YPD, plated in YPD at a cell density of 200 cells and grown overnight for two days at 30 ºC. The presence of the i-5L chromosome was assessed in 10 colonies per plate via PCR using the primers AB1015 and AB1019 (**Table 2.6**) and the percentage of loss was determined. Actin PCR was used as a positive control (primers AB800 and AB801 [**Table 2.6**]) A positive control for chromosome 5 annealing was also conducted (primers AB1015 and AB1016 [**Table 2.6**]) The experiment was conducted in 3 replicates, for an analysis a total of 210 colonies. The absence of the chromosome i-5L was further confirmed in 3 of the negative strains by Southern Blotting.

2.8 Adaptive laboratory evolution (ALE)

A single colony of the *S. stipitis* strain NRRL Y-7124 was grown overnight in 5 ml of YPD at 30 ºC, plated in YPD at a cell density of 100 and grown 48 hours at 30 ºC. 36 selected as starting colonies were re-straked in two SC-Mix plates and grown at 30 ºC and 37 ºC, respectively. Moreover, a sample of each colony was frozen in glycerol at -80 ºC as a representation of the karyotype at the starting point. The strains were then streaked into new SC-Mix plates when stationary growth was reached (daily for controls at 30 ºC and every two days for the first two weeks at 37 ºC, and daily after that) for a total of 8 weeks. The karyotype variability of the colonies was assessed by CHEF electrophoresis (See section 2.5)

2.9 Strain phenotyping

2.9.1 Growth evaluation

To investigate growth on different carbon source, *S. stipitis* strains were grown overnight in 5 ml of SC-Glucose, SC-Xylose or SC-Mixture at 30 ºC with agitation of 200 rpm. The growth was also evaluated in the presence of different inhibitors commonly found in lignocellulose hydrolysates.

The overnight culture was used to inoculate a 96 well culture plates (Cellstar®, #655180) at 60 cells/µl in 100 µl. The cells were then grown in a BMG Labtech SPECTROstar nano plate reader at 30 ºC with double orbital agitation of 400 rpm. OD_{600} was measured for a total of 48 hours.

Growth rate, maximum OD_{600} and lag time were determined as growth parameters. The growth rate was calculated as the difference between the Napierian logarithm of the

biomass at two time points during the exponential phase of microbial growth, and it was measured in hours^{-1}:

$$\frac{ln(X2) - ln\ (X1)}{t2 - t1}$$

Where: (i) X1 is the biomass concentration (OD_{600}) at time point one (t1) (ii) X2 is the biomass concentration (OD_{600}) at time point two (t2).

The maximum OD was determined with the MAX() function present in Excel (Microsoft®), and measured in OD units. The lag time was determined visually as the time in which the exponential growth starts, in minutes. The results were plotted using Excel or Rstudio (R version 4.0.2) and outliers removed from all the analysis. Experiments were performed in 3 technical and 3 biological replicates.

2.9.2 Agar invasion

The analysis was conducted following the protocol described by Hope and Dunham (Hope and Dunham 2014). Strains grown overnight in SC-Mix were used to inoculate fresh media at an OD of 0.1 and were grown until mid-exponential phase (OD 0.7-1.0, (Eppendorf BioSpectometer® basic, #6135000025)). 200 µl of the culture were used to spot SC-Mix agar plates with a replica plater (Replica plater for 96 well plates (8x6) Sigma Aldrich, #R2383). 3 biological replicates were included in each plate and three plates were inoculated as technical replicates. A strain of *S. cerevisiae* (BY4741) and *C. albicans (SC5314)* were used respectively as negative and positive control of agar invasion. The reference strain Y-11545 was included in all the plates to allow better comparisons. The plates were grown for 5 days at 30 ºC and 37 ºC, after which, the colonies were washed off the surface under a stream of distilled water. The spots were photographed before and after washing. The washed plates were then grown for additional 24 hours and then washed and photographed again to observe additional growth from cells trapped in the agar.

The images were converted to black and white analysed in the image processing software ImageJ v 1.51 j8. To quantify the invasion for each spot, the mean grey intensity of the spot and the surrounding background were determined. To correct the background (B_c), the following formula was used:

$$B_c = \frac{A_t * I_t - A_s * I_s}{A_t - A_s}$$

Where: (i) A_t is the area surrounding the three spots, (ii) I_t is the mean grey intensity of the region selected as A_t, (iii) A_s is the area of the spot, (iv) I_s the mean grey intensity of the spot.

Hence, the quantitative agar invasion (Q_a) will be:

$$Q_a = I_s - B_c$$

For the strains where the background correction yielded a negative number, the value was corrected to zero.

To correct agar invasion respect to the growth of the strains on the surface, the relation between mean grey intensity of the spot before and after washing was determined, following the formula proposed by Zupan and Raspor (Zupan and Raspor 2008):

$$R_i = \frac{Qa}{I_{sp}} * 100$$

Where: (i) Ri is the relative invasion (ii) I_{sp} is the mean grey intensity before the colony is washed.

2.9.3 Sedimentation assay

The analysis was conducted according to the protocol proposed by Guebel and Nudel (Guebel and Nudel 1994), with modifications. Strains were grown overnight in SC-Mix and the concentration of the cultures was determined spectrophotometrically at OD_{600} (Eppendorf BioSpectometer® basic, #6135000025). The volume of overnight culture needed to obtain an OD_{600} of 1.0 was calculated and added to milli-Q water to form a 1 ml solution. The solution was then transferred to a cuvette and the OD was measured every hour, with no agitation between measurements, for 4 hours. The percentage of settling was calculated according to the equation:

$$Sedimentation\ (\%) = 100 - (\frac{A_{t_x}}{A_{t_0}} * 100)$$

Where: (i) A_{tx} is the absorbance recorded at time x at OD_{600} (ii) A_{t0} is the absorbance recorded at time 0 at OD_{600}.

The experiment was conducted in three biological replicates.

2.9.4 Statistical analysis

The growth parameters were compared by ANOVA tests using R studio (R version 4.0.2). The equality of variances presumption to be able to use ANOVA in statistical comparisons was checked by the Levene's test for equality of variances, whereas the normality of the data was checked by Shapiro-Wilk normality test. When both presumptions

were obeyed the Tukey's honest significant difference test was used to determine were the differences stated by the ANOVA test lied.

On the other hand, when the variances were not statistically the same, the one-way test was used to indicate significance between the phenotypes assessed. When significant, the pairwise testing was used to determine were those differenced lied. In case of equal variances, but no normal distribution of the data, the Kruskal-Wallis rank sum test was used to indicate statistical differences.

When parametric variables were compared the Pearson correlation test was used, whereas the chi square test was used to test independence of datasets.

A p-value lower than 0.05 was considered significant for all the statistic tests.

Table 2.6. Primers used in this study.

Primer ID	Name	Sequence
AB796	its1	TCCGTAGGTGAACCTGCGG
AB797	its4	TCCTCCGCTTATTGATATGC
AB798	NL1	GCATATCAATAAGCGGAGGAAAAG
AB799	NL-4	GGTCCGTGTTTCAAGACGG
AB800	ActF	TTCCGGTATGTGTAAGGCCG
AB801	ActR	TTCATTGGGGCTTCGGTCAA
AB966	589_C2_Ups	GGAAGCTGGGTTAGTCGTGT
AB967	589_C2_Down	ACACGGAAAGTTCTGCCCAT
AB968	589_C1_Ups	TTAGTCTATATGCGGAGAGCCG
AB969	589_C1_Down	GAAAAGAGGTCTCCTCGGGC
AB1015	MChr_check_8	GCAAGATGGACCCGGACTCA
AB1016	MChr_check_9	GCAAGCTTATGATGAAGTTTTGCGA
AB1019	MChr_check_11	AAGATCCATACCCAATCGTG
AB1028	South_Trans_C5_F	CAACTTCAAACACCGGCTCG
AB1029	South_Trans_C5_R	CTGGTGTCGACGGTAAACCA

Table 2.1. Strains used for this project.

Collection code	Habitat
ATCC 58784 CBS 5776 NRRL Y-11544	Insect (IN) larva on fruit tree, Rhône, Lyon, France.
ATCC 58376 CBS 5773 NRRL Y-7124	Insect (IN) larva on fruit tree, Rhône, Lyon, France.
ATCC 62970 CBS 5774 NRRL Y-11542	Insect (IN) larva on fruit tree, Rhône, Lyon, France.
ATCC 62971 CBS 5775 NRRL Y-11543	Insect (IN) larva on fruit tree, Rhône, Lyon, France.
CBS 7126 NRRL Y-17104	Industrial fermentations (IFM), from xylose fermentation.
NRRL YB-3756	Pine and coniferous trees (PTR), gymnosperms, dead pine tree, Gainesville, Florida, USA.
NRRL Y-27547	Insect (IN), beetle, Odontotaenius disjunctus, Passalidae, Burke Co. Shell Bluff, Georgia, USA.
NRRL Y-27548	Insect (IN), beetle, Odontotaenius disjunctus, Passalidae, Lake Herrick Park, Georgia, USA
NRRL Y-27549	Insect (IN), beetle, Passalidae, Orangeburg County, South Carolina, USA.
NRRL Y-27550	Insect (IN), beetle, Odontotaenius disjunctus, Passalidae, Burden, Baton Rouge, Louisiana, USA.
NRRL YB-1611	Hard wood trees (HTR), angiosperms, shagbark hickory, Peoria, Illinois, USA.
NRRL Y-12759	Soil or rock (SL), forest soil, Georgia, USA.
NRRL YB-3713	Frass or Insect tunnels (FR), frass on dead oak log, Gainesville, Florida, USA.
NRRL Y-27552	Insect (IN), beetle, Verres sternbergianus, Passalidae, Barro Colorado Island, Panama.
NRRL Y-27535	Insect (IN), gut of a Passalid beetle, Kansas, USA.
NRRL Y-8209	Tree, unknown type (TR), tree trunk, near New Orleans, Louisiana, USA.
NRRL Y-8271	Frass or insect tunnels (FR), frass, near New Orleans, Louisiana, USA.
NRRL Y-11545	Unknown.
NRRL Y-17100	Insect (IN), insect larvae on fruit tree, Rhône, Lyon, France.
NRRL Y-27551	Insect (IN), beetle, Odontotaenius disjunctus, Passalidae, Duches Dr. Park, Baton Rouge, Louisiana, USA.
NRRL Y-27555	Insect (IN), beetle, Odontotaenius disjunctus, Passalidae, Oxford, Pennsylvania, USA.
NRRL YB-1337	Tree, unknown type (TR), rotted log, Wohlwend Farm, Marion, Illinois, USA.
NRRL YB-1762	Frass or insect tunnels (FR), frass, rotten logs, Brownfield Woods, Illinois, USA.
NRRL YB-2051	Frass or insect tunnels (FR), frass, dead black oak.
NRRL YB-3619	Frass or insect tunnels (FR), frass, American elm.
ATCC 58376, CBS 5773 NRRL Y-7124	Insect (IN) larvae on fruit tree, Rhône, Lyon, France.
NRRL Y-27553	Insect (IN), beetle, Verres sternbergianus, Passalidae, Barro Colorado Island, Panama.
NRRL Y-50871 NRRL Y-50859 NRRL Y-50862 NRRL Y-50861	Strains adapted from Y-7124 to industrially promising media by Slininger *et al*. (2015) Gently ceded by Patricia Slininger. NRRL, Peoria, Illinois, USA.

Chapter 3

Repetitive elements and plasticity in the genome of natural isolates of *Scheffersomyces stipitis*

3.1 INTRODUCTION

Living organisms are importantly defined by their genomes. Control of the gene content and expression can determine the success in the adaptation to potential stresses that challenge survival. Therefore, the potential of acquisition of modifications that enhance survival in challenging conditions are important for cell subsistence and evolution. This alterable nature in the genetic information of organisms is known as genome plasticity.

Genome plasticity has been widely studied in microorganisms, mainly in pathogenic bacteria and yeasts, given the importance of understanding the modifications that their genomes suffer when treatments are carried out (Bennett 2004) (Legrand *et al.* 2019; Tsushima *et al.* 2019). This studies have been conducted even before whole genome sequencing was available, and were based in the determination of chromosome rearrangements, aneuploidies or copy number variations (polyploidy) mainly by karyotyping techniques, such as pulse field gel electrophoresis (PFGE) (Rustchenko-Bulgac 1991; Suzuki *et al.* 1982; Chibana 2000).

Special emphasis has been dedicated in the study of the genome plasticity of the opportunistic human pathogen *C. albicans*, since it is the fourth most common infection treated in hospitals (Fisher *et al.* 2018), and can be fatal in ~50% of cases (Antinori *et al.* 2016). This yeast belongs to the CTG-clade, a group of yeast characterised by an alternative genetic code in which the CUG codon codifies for a serine instead of a leucine (Krassowski *et al.* 2018).

Different DNA repeats have been linked with genome instability in *C. albicans,* which might condition the wide plasticity observed in clinical isolates. Some of these elements include long repetitive elements, long terminal repeats and transposons, short tandem repeats, or long inverted repeats (Buscaino 2019). These effects have also been observed in other pathogenic yeasts of the clade, such as *C. glabrata,* where repetitive genes that codify for cell wall proteins show wide plasticity, which also correlates with the high genomic diversity observed in different isolates (Carreté *et al.* 2018), or *C. dubliniensis,* which exhibits similar repeats distribution but greater translocation frequency when compared to *C. albicans* (Magee *et al.* 2008). However, although genetic and phenotypic diversity has been observed in non-pathogenic species of the clade, little is known about the genetic reasons behind it (Jacques *et al.* 2010; Piombo *et al.* 2018).

Therefore, the aim of this chapter was to test and understand whether the genome of *S. stipitis*, a non-pathogenic yeast present in the CTG clade, but widely used for bioethanol fermentation research, is also plastic. Previous studies conducted with 4 natural isolates of *S. stipitis* isolated from insect larvae (ATCC 58784, ATCC 58376,

ATCC 62970, ATCC 62971) and 1 from a xylose fermenter (Y-17104) had shown small variations at chromosome organization by transverse-alternating field gel electrophoresis (TAFE) (Passoth *et al.* 1992). In that study, Passoth *et al.* identified a total of six chromosomes (bands) for all five strains, and the banding profiles could be classified in four different groups. However, no further analyses have been conducted to explain the nature of that variability or to determine if strains from different habitats also exhibit plasticity.

To investigate this, this chapter first analyses by CHEF electrophoresis the karyotype of 27 natural isolates of *S. stipitis* obtained from the Agricultural Research Service (ARS) Collection, run by the Northern Regional Research Laboratory (NRRL) (Peoria, Illinois, USA), and from the National Collection of Yeast Cultures (NCYC) (Norwich, United Kingdom). The strains were collected in four different countries: United States of America (USA), France, Panama and The Netherlands, and from six different habitats (insect larvae, insect gastrointestinal tract, soil, frass, trees and from a xylose fermentation reactor) (See **Table 2.1** for full details).

Subsequently, the genome sequence of two strains that exhibited different karyotype is compared, with special emphasis in repetitive regions. The strain NRRL Y-11545 had been previously sequenced by shotgun sequencing and published by Jeffries *et al.* (Jeffries *et al.* 2007), whereas the strain Y-7124 has been sequenced in this study by hybrid TGS, involving both ONT and Illumina sequencing.

3.2 RESULTS

3.2.1 Identification of *S. stipitis* strains

The first step conducted in this project was the genetic identification of the 27 strains present in the collection. To do so, the regions spanning the internal transcribed spacers ITS1 and ITS2 and the 5.8S gene (5.8S-ITS rDNA) and the sequence of the D1/D2 domain of the 26S rDNA gene were amplified and sequenced, as recommended by Villa-Carvajal *et al*. (Villa-Carvajal, Querol and Belloch 2006). As a control, a reference strain was included (NRRL Y-11545), since its genome has been fully sequenced and a high quality assembly is available (Jeffries *et al.* 2007).

All the strains present in the collection were identified as *S. stipitis.* In agreement with the results obtained by Villa-Carvajal *et al*., Sanger sequencing of the D1/D2 domain of the 26S rDNA showed better success in the identification of all 27 natural isolates of

S. stipitis, whereas the region 5.8S-ITS rRNA gene failed to correctly identify three natural isolates (NRRL Y-27535 was identified as *Enteroramus dimorphus* and NRRL YB-1337 and NRRL Y-27553 as *Scheffersomyces illinoinensis*) and to determine the exact species of 5 natural isolates (NRRL Y-17104, NRRL Y-27550, NRRL Y-27551, NRRL Y-1762 and NRRL Y-7124 were identified only as species belonging to the genus *Scheffersomyces*) (**Table 3.1**).

Table 3.1. Identification of *S. stipitis* natural isolates by Sanger sequencing of both D1/D2 domain of 26S rDNA gene and 5.8S-ITS rRNA gene.

Code	Blast result: D1/D2 domain of 26S rDNA gene	% identity	E-value	Blast result: 5.8S-ITS rRNA gene	% identity	E-value
ATCC 58784	*Scheffersomyces stipitis*	99.47	0.0	*Scheffersomyces stipitis*	99.83	0.0
ATCC 58376	*S. stipitis*	99.14	0.0	*S. stipitis*	100	0.0
ATCC 62970	*S. stipitis*	98.63	0.0	*S. stipitis*	99.48	0.0
ATCC 62971	*S. stipitis*	99.65	0.0	*S. stipitis*	99.66	0.0
NRRL Y-17104	*S. stipitis*	99.65	0.0	*Scheffersomyces sp*	99.48	0.0
NRRL YB-3756	*S. stipitis*	99.65	0.0	*S. stipitis*	99.66	0.0
NRRL Y-27547	*S. stipitis*	99.82	0.0	*S. stipitis*	99.83	0.0
NRRL Y-27548	*S. stipitis*	99.65	0.0	*S. stipitis*	99.83	0.0
NRRL Y-27549	*S. stipitis*	99.47	0.0	*S. stipitis*	99.66	0.0
NRRL Y-27550	*S. stipitis*	99.82	0.0	*Scheffersomyces sp*	100	0.0
NRRL YB-1611	*S. stipitis*	99.82	0.0	*S. stipitis*	100	0.0
NRRL Y-12759	*S. stipitis*	99.65	0.0	*S. stipitis*	99.66	0.0
NRRL YB-3713	*S. stipitis*	99.13	0.0	*S. stipitis*	99.83	0.0
NRRL Y-27552	*S. stipitis*	99.65	0.0	*S. stipitis*	99.47	0.0
NRRL Y-27535	*S. stipitis*	99.65	0.0	*Enteroramus dimorphus*	100	0.0
NRRL Y-8209	*S. stipitis*	99.82	0.0	*S. stipitis*	100	0.0
NRRL Y-8271	*S. stipitis*	99.47	0.0	*S. stipitis*	99.83	0.0
NRRL Y-11545	*S. stipitis*	99.48	0.0	*S. stipitis*	100	0.0
NRRL Y-17100	*S. stipitis*	99.65	0.0	*S. stipitis*	99.66	0.0
NRRL Y-27551	*S. stipitis*	99.47	0.0	*Scheffersomyces sp*	99.83	0.0
NRRL Y-27555	*S. stipitis*	99.29	0.0	*S. stipitis*	100	0.0
NRRL YB-1337	*S. stipitis*	99.82	0.0	*S. illinoinensis*	99.33	0.0
NRRL YB-1762	*S. stipitis*	99.65	0.0	*Scheffersomyces sp*	100	0.0
NRRL YB-2051	*S. stipitis*	99.82	0.0	*S. stipitis*	100	0.0
NRRL YB-3619	*S. stipitis*	99.31	0.0	*S. stipitis*	99.33	0.0
NRRL Y-7124	*S. stipitis*	99.65	0.0	*Scheffersomyces sp*	100	0.0
NRRL Y-27553	*S. stipitis*	99.82	0.0	*S. illinoinensis*	99.65	0.0

Despite of this, the ITS rRNA gene is commonly used for the phylogenetic analysis of different species. Therefore, the sequences obtained for the natural isolates of *S. stipitis* were aligned and a maximum likelihood tree was constructed to understand how closely related each specie is according to the 5.8S-ITS gene (**Figure 3.1**). This result suggests that the strains Y-11545 and Y-7124, used for strain genome comparison

in this project, show early divergence in the strain evolution history of *S. stipitis.* However, it is important to highlight that this phylogenetic tree is constructed only with one gene that is strongly conserved with low variability, and therefore its resolution might not be optimal for strain differentiation.

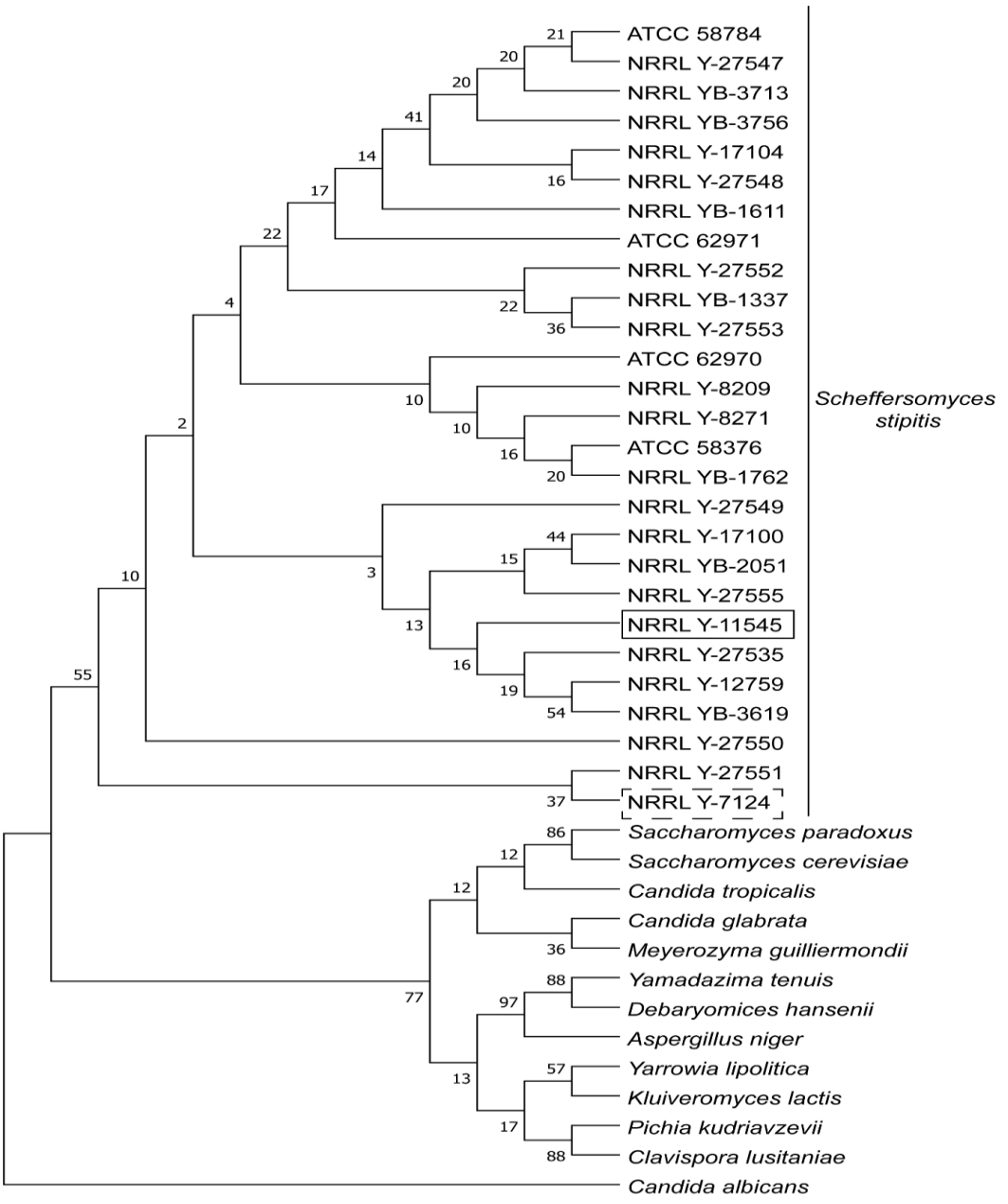


Figure 3.1. Maximum likelihood phylogenetic tree of the 5.8S-ITS rRNA gene of the different natural isolates of *S. stipitis* studied in this project. Several species from the Saccharomycetales order were included for comparison purposes. The black box indicates the natural isolate whose genome was previously sequenced (Jeffries *et al.* (2007)). Dashed box indicates the other natural isolate sequenced in this project.

3.2.2 Genome structure diversity in *S. stipitis* natural isolates

To assess whether genome diversity is common across *S. stipitis* isolates, the genome structure of 27 natural isolates was analysed by CHEF electrophoresis, a technique that allows chromosome separation according to size (Schwartz and Cantor 1984; Chu, Vollrath and Davis 1986).

Genome-wide Illumina sequencing of the strain NRRL Y-11545 has indicated that *S. stipitis* is a haploid yeast, with a genome organised in eight chromosomes raging in size from 3.5 to 0.97 Mbp (Jeffries *et al.* 2007). Since this is the only strain sequenced and assembled to chromosome level to date, it is used as a reference strain for the genetic studies in *S. stipitis*. The karyotype of the strain shows the presence of only six bands, instead of the eight expected (**Figure 3.2**). That can be explained by the fact that chromosomes three and four and five and six have the same size (1.8 Mbp and 1.7 Mbp respectively), and therefore they co-migrate in the gel and cannot be resolved. This hypothesis is also supported by a higher brightness in the bands corresponding to those chromosomes, indicating the presence of higher concentration of DNA.

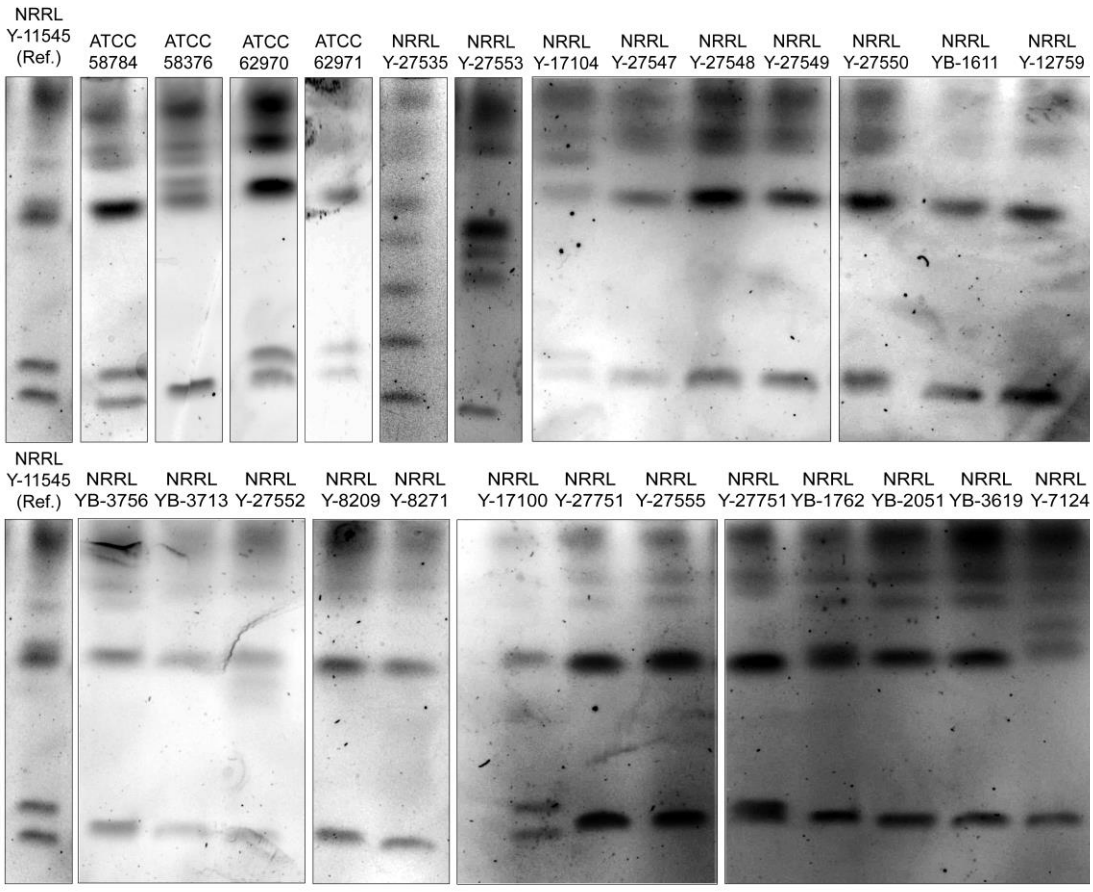


Figure 3.2. CHEF Electrophoresis of *S. stipitis* natural isolates. Running conditions: 6 V/cm, switch time: 60-120s for 12 hours. 4.5 V/cm, switch time: 120-300s switch time for 12 hours at 4.5 V/cm. Staining in ethidium bromide (0.5 µg/ml).

The CHEF electrophoresis results (**Figure 3.2**) demonstrate a wide genome diversity associated to *S. stipitis*, since clear differences are observed across the chromosome banding pattern of the natural isolates studied in this project. The number of bands varies from five to seven.

Despite the wide variability observed, a discrete number of genome organisations is detected, and the strains can be clustered in eight different groups (G1 to G8, **Figure 3.3**).

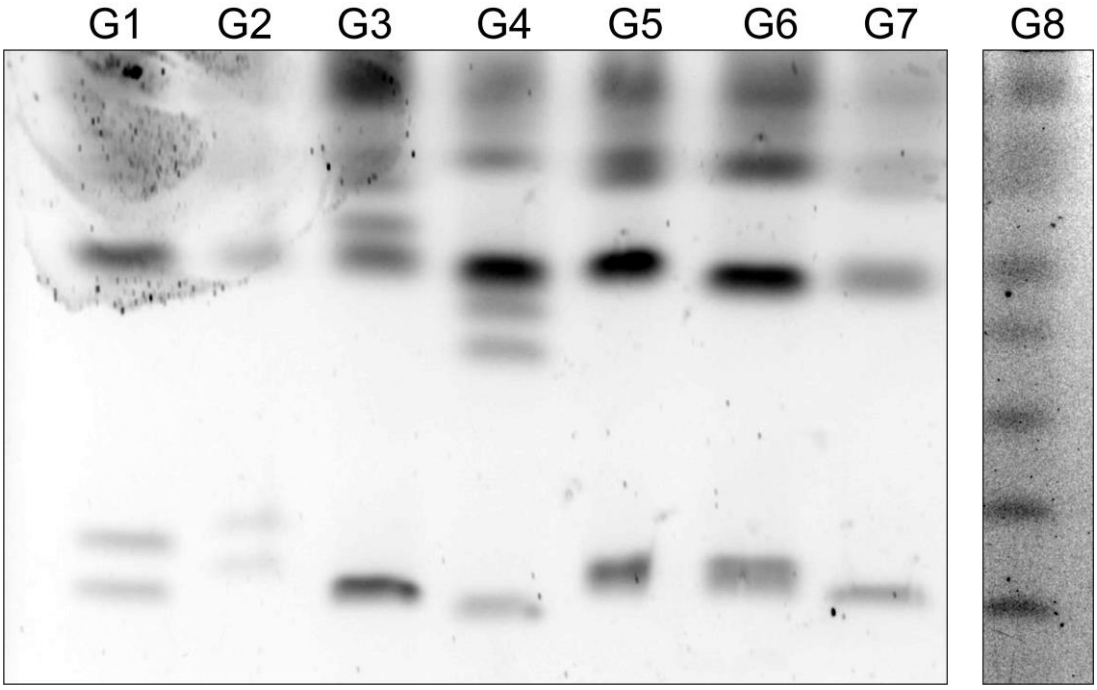


Figure 3.3. Classification of the different CHEF Electrophoresis observed for *S. stipitis* natural isolates. Running conditions: 60-120s switching time for 12 hours at 6 V/cm; switching time of 120-300s for 12 hours at 4.5 V/cm. Staining in ethidium bromide (0.5 µg/ml).

The most common genome structure is the exhibited by the strains in G5 (n=15) (NRRL YB-3756, NRRL Y-27547, NRRL Y-27548, NRRL Y-27549, NRRL Y-27550, NRRL YB-1611, NRRL Y-12759, NRRL YB-3713, NRRL Y-27552, Y-8209, Y-8271, NRRL Y-27551, NRRL Y-27555, NRRL YB-2051 and NRRL YB-3619). Five groups are formed by only one strain: G2 (ATCC 62971), G4 (NRRL Y-27553), G6 (NRRL YB-1337), G7 (NRRL YB-1762) and G8 (NRRL Y-27535). G1 is formed by five strains: ATCC 58784, ATCC 62970, NRRL Y-17104., NRRL Y-11545 and NRRL Y-17100, and G3 by two, ATCC 58376 and NRRL Y-7124. Notably, the classification in the different karyotypes is not consistent with the classification in the phylogenetic analysis the 5.8S-ITS gene. This might be related with the low accuracy previously reported for single-gene phylogenetic trees (Castresana 2007).

Therefore, the 27 natural isolates studied in this project exhibit a total of 8 different karyotypes. These findings can be explained by mainly two theories: (i) Natural

isolates belonging to the same group have a common origin or evolution, or (ii) There is a discrete number of genome organizations that are stable and viable. However, despite of the wide variation described, it should be noted that small genomic differences are not captured by CHEF electrophoresis, and that therefore this group classification might not be optimal.

3.2.3 Genome organisation of the *S. stipitis* Y-7124 natural isolate

The next step in this project was studying the nature of the wide genome plasticity observed in the different natural isolates. To address this, the full genome of one natural isolate with structural variations respect to the reference strain was sequenced. The strain NRRL Y-7124 was selected since it is a commonly used strain in fermentation studies (J.P. Delgenes, Moletta and Navarro 1988; Mussatto *et al.* 2012; Slininger *et al.* 2015; Koti *et al.* 2016; Corro-Herrera *et al.* 2018).

The genome was sequenced using hybrid TGS approach, combining ONT and Ilumina sequencing, which allowed the generation of a final new assembly with a total coverage of 186.88x over 15.69 Mb, arranged in 11 contigs. The identification of unique centromeric sequences (present once in each chromosome) guided the final assembly into complete chromosomes (**Tables S1 and S2**).

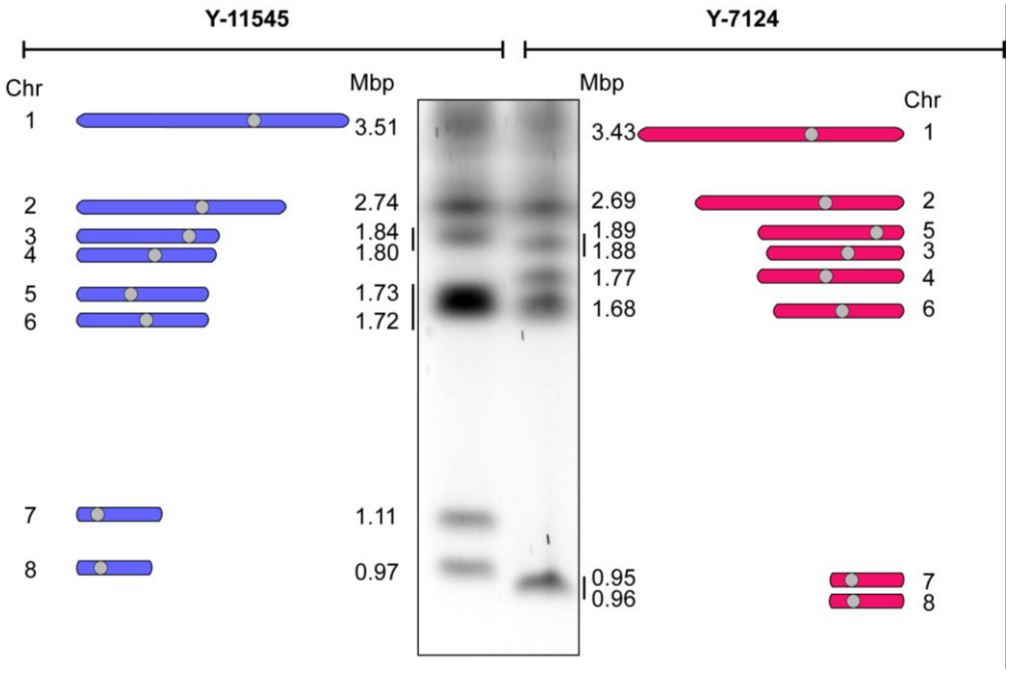


Figure 3.4. Comparison of the genome organization of *S. stipitis* strains NRRL Y-7124 and NRRL Y-11545. CHEF Electrophoresis running conditions: 60-120s switching time for 12 hours at 6 V/cm; switching time of 120-300s for 12 hours at 4.5 V/cm. Staining in ethidium bromide (0.5 µg/ml).

Comparison of the Y-7124 and the Y-11545 (reference) genomes demonstrated that the two natural isolates have genomes of similar size (Y-7124: 15.26 Mbp; Y-11545: 15.44 Mbp) organised both in 8 chromosomes. However, individual chromosomes have a different size and chromosomal organisation, which explains the differences previously observed in the karyotype (**Figure 3.4**).

There is a total of 50,495 SNPs between the two natural isolates (**Table 3.2**), with 1 variant every 306 bases. Despite of the high number of SNPs, 16,294 (74.25%) of them are synonymous (no changes in the protein sequence), 5,622 (25.62%) are missense (changes in the amino acid sequence of the protein) and only 28 (0.13%) are nonsense (appearance of a premature stop codon) (**Table 3.3**).

Table 3.2. Summary of the SNPs observed between the natural isolates of *S. stipitis* NRRL Y-11545 and NRRL Y-7124.

Number of variants	50,495
Variant rate	1 variant / 306 bases
Transitions	30,655
Transversions	19,725
Region	
Exon	21,904 (43.38%)
Intergenic	25,680 (50.86%)
Other	2,911(5.85%)
Effects	
Missense	5,622 (25.62%)
Nonsense	28 (0.13%)
Silent (synonymous)	16,294 (74.25%)

The first annotation in the genome of the strain Y-11545 predicted a total of 5,841 genes, using the Joint Genome Institute (JGI) Annotation Pipeline (Jeffries *et al.* 2007). However, Maguire *et al*. (Maguire *et al.* 2013) demonstrated errors in the this prediction and updated the total number of codifying genes to 6,026 using SearchDOGS (OhÉigeartaigh *et al.* 2011). Our study predicted a total of 6,330 proteins in the strain Y-7124 using FunGAP v.1.1.0 (Min, Grigoriev and Choi 2017). This important difference observed in the gene content between the two strains (304 predicted genes) is probably due to misidentifications in the gene prediction of the strain Y-7124, since a number genes identified in Y-11545 were manually identified in Y-7124 as several split ORFs, whose combination would yield the complete gene in the reference strain. However, the theory of differences related to the use of different programs for gene prediction must not be discarded.

Gene conservation between the two strains was studied by determination of orthologous genes with the web server OrthoVenn2 (Xu *et al.* 2019). This program classifies all the protein models in different clusters, where each cluster will consist in

orthologs or paralogs from the species studied. The total 5,841 proteins predicted in the original annotation of the strain Y-11545 were classified in 5,639 clusters, whereas the total 6,330 proteins predicted in this study for the strain Y-7124 were classified in 5,646 (**Figure 3.5**). A total of 5,630 clusters were present in the two strains, which represents 11,378 proteins (5,684 (or 49.96%) in Y-11545 and 5,694 (or 50.04%) in Y-7124). According to the program, the strain Y-11545 presents 9 unique clusters, which include 37 proteins, whereas in Y-7124 there are 43 unique proteins divided in 16 clusters (**Table 3.4 and 3.5**), although some of the proteins identified by OrthoVenn2 as unique clusters in each strain has been detected in the other (Table H and I). Moreover, 114 proteins in the strain Y-11545 were not forming any cluster (singleton), whereas 599 singletons were detected in Y-7124. Therefore, despite the high number of SNPs observed between the two strains and that are relatively distant phylogenetically according to the 5.8S-ITS gene (**Figure 3.1**), the gene content is similar.

Table 3.3. Non-sense SNPs detected in the study between the natural isolates Y-11545 and Y-7124. Chr: Chromosome. N.A: Non-annotated. [a]: Annotated in this study.

Protein Id	Chr	Gene	Description
PICST_28603	1	*PICST_28603*	N.A
PICST_28677	1	*HWP1*	Hyphal wall protein similar to FLO1
PICST_29191	1	*PICST_29191*	N.A
PICST_29289	1	*PICST_29289*	N.A
PICST_37563	1	*MRP10*	Mitochondrial/chloroplast ribosomal protein L15/L10 Translation
PICST_37697	1	*ADD6*	Aryl-alcohol dehydrogenases
PICST_52332	1	*PICST_52332*	ADP ribosylation factor
PICST_80517	1	*HXT2.4*	Probable hexose transporter
PICST_29938	2	*PICST_29938*	N.A
PICST_56002	2	*PICST_56002*	N.A
PICST_56221	2	*PICST_56221*	N.A
PICST_40512	2	*GHF18*	Glycosyl hydrolase family 18; possible chitin bindin
PICST_81787	2	*KRE9*	Cell wall synthesis protein KRE9 precursor
PICST_30958	3	*PICST_30958*	N.A
PICST_42867	3	*FRE1.4*	Ferric reductase transmembrane component
PICST_31037	3	*PICST_31037*	OPT transporter protein
PICST_59158	4	*PICST_59158*	N.A
PICST_32113	5	*PICST_32113*	LINEA1[a]
PICST_32479	5	*MED11*	Mediator subunit of RNA polymerase II
PICST_47013	5	*MSR2.2*	Mitochondrial protein with homology to MRS2
PICST_73033	6	*PICST_73033*	N.A.
PICST_63741	7	*PICST_63741*	OPT transporter protein
PICST_33744	8	*SPF1.3*	SPF1 P-type ATPase
PICST_33820	8	*QDR22*	Multidrug resistance transporter
PICST_33956	8	*VPS70*	Membrane protein involved in vacuolar protein sorting
PICST_64492	8	*HYR5.5*	Hyphally regulated cell wall protein

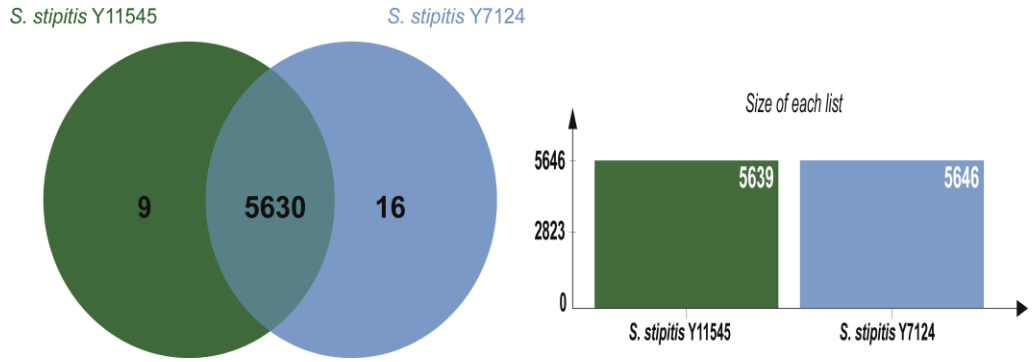


Figure 3.5. Venn Diagram with the number of clusters determined for each *S. stipitis* strains NRRL Y-11545 and NRRL Y-7124.

Table 3.4. Unique clusters identified in the strain *S. stipitis* NRRL Y-11545. [a]: Original annotation by Jeffries *et al.* (2007). Abbreviations: N.A: Not annotated. Y: Yes.

Cluster ID	Proteins (n)	GO Annotation (OrthoVenn2)	Genes [a]	Genes description [a]	Present in Y-7124
1	11	N/A	*ADD2.4, VCC3.1, ADD2.3, ADD3.1, ADD3.3, ADD2.2 PICST_33214, VCC3.3, ADD2.1 ADD3.2, VCC3.5*	ATP-dependent DNA helicases.	Y
2	10	N/A	*PSR2.2, PSR2.4 PSR2, PSR2.1, PICST_28618 PICST_33381, MSR2.3, MUC1.10*	Cytochrome c binding site.	Y
11	6	Trichloro-p-hydroquinone reductive dehalogenase activity	*OYE2.1, OYE2.5 OYE2.6, OYE2.7 OYE2.8, OYE2.9*	NADPH dehydrogenase (OYE, Old yellow enzymes)	Y
15	5	Vesicle-mediated transport	*SSA2.1, SSA2.2 HSP70.1, HSP70.3 KAR2*	Heat shock proteins and molecular chaperones.	Y
53	3	Pathogenesis	*HYR6.6, HYR6.3, HYR6.4*	Hyphally regulated cell wall protein	Y
1947	2	Sphingolipid biosynthetic process	*INPT2 IPT1*	Inositolphosphotransferase Putative inositol phosphoesterase	
1948	2	N/A	*PICST_33382, PICST_33387*	N.A.	
1949	2	Pathogenesis	*HYR2.1, HYR2.1*	Hyphally regulated cell wall protein	Y
1950	2	N/A	*PSR1.1, PSR1.2*	N.A.	

Table 3.5 Unique clusters identified in the strain *S. stipitis* NRRL Y-7124. [a]: Genes annotated in this study. [b]: Genes identified by BLASTN. N.A: Non-annotated. Y: Yes.

Cluster ID	Proteins (n)	GO Annotation (OrthoVenn2)	Genes[a]	Annotation	Present in Y-11545
36	4	N/A	*gene_02013*	No similarity.	
			gene_04329	No similarity.	
			gene_05074	No similarity.	
			gene_05244	No similarity.	
39	3	N/A	*gene_04157*	LINEA1[a]	Y
			gene_03689	LINEA1[a]	
			gene_05497	LINEA1[a]	
56	3	Transcription DNA-templated	*gene_02673*	NUPAVs (Nuclear sequences of plasmid and viral origin) in *S. stipitis* NRRL Y-11545	Y
			gene_02672		
			gene_02674		
57	3	generation of catalytic spliceosome for first transesterification step	*gene_00356*	N.A (P-loop containing nucleoside triphosphate hydrolase protein in *Suhomyces tanzawaensis*)[b]	Y
			gene_00025	N.A. (Helicase conserved C-terminal domain-containing protein in *Debaryomyces* fabryi)[b]	
			gene_02993	N.A. (pre-mRNA-splicing factor ATP-dependent RNA helicase PRP16 in *C. albicans*)[b]	
68	2	Integral component of membrane	*gene_04598*	N.A.	
			gene_03430	N.A.	
69	2	N/A	*gene_06297*	*PICST_47120* in *S. stipitis* NRRL Y-11545[b]	Y
			gene_05915	*PICST_47120* in *S. stipitis* NRRL Y-11545[b]	
70	2	N/A	*gene_02692*	N.A.	
			gene_05691	N.A.	
71	2	N/A	*gene_02698*	N.A.	
			gene_02699	N.A.	
72	2	Termination of RNA polymerase II transcription	*gene_01941*	N.A (5'-3' exoribonuclease 2 in *Debaryomyces* fabryi)[b]	
			gene_01241	N.A. (5'-3' exoribonuclease 1 in *C.* albicans)[b]	
73	2	Regulation of Rho protein signal transduction	*gene_06323*	N.A.	
			gene_05939	N.A.	
74	2	N/A	*gene_02086*	N.A.	
			gene_05166	N.A.	
1951	2	Cytochrome c oxidase activity	*gene_06330*	N.A. (Cytochrome C oxidase subunit II in *Debaryomyces fabryi*)[b]	
			gene_06328	N.A. (Cytochrome C oxidase subunit II in *Debaryomyces fabryi*)[b]	
1952	2	Telomere maintenance	*gene_06247*	N.A. (ATP-DNA helicases in *Spathaspora* sp)[b]	Y
			gene_05682	N.A. (ATP-DNA helicases in *Candida* parapsilosis)[b]	
1953	2	RNA-directed DNA polymerase activity	*gene_01508*	ORF2 Zorro 3 TE[a]	Y
			gene_05599	ORF2 Zorro 3 TE[a]	
1954	2	N/A	*gene_05101*	No similarity.	
			gene_02036	No similarity.	
1955	2	Regulation of translation termination	*gene_02801*	N. A. (*PICST_73544* in *S. stipitis* NRRL Y-11545)	Y
			gene_00542	N.A. (P-loop containing nucleoside triphosphate hydrolase protein in *Suhomyces tanzawaensis*)[b]	

3.2.4 The contribution of repetitive elements to genome plasticity

Repetitive elements are major contributors to genome plasticity as repeats can undergo both inter and intra-locus recombination events, leading to chromosome translocations, deletions, duplications and chromosomal fusions (Freire-Benéitez, Price, *et al.* 2016; Dunn and Anderson 2019; Robert T. Todd *et al.* 2019). Accordingly, structural variation often originates from repetitive elements such as transposons (Alonge *et al.* 2020).

A comprehensive survey of the major classes of repetitive elements decorating the *S. stipitis* genome has not been performed yet. Given the central role of repetitive elements in regulating genome plasticity, this study sought to classify all the major classes of repetitive elements within the available *S. stipitis* genomes (NRRL Y-11545 and NRRL Y-7124). To this end, long sequences (>100 nucleotides) present more than once in the genome were identified by aligning the genomes sequenced to themselves by using BLAST-N. The genomic position of these repeats was manually verified using IGV/SNAPGene, and clustered repeats were combined. This analysis identified different types of intra-chromosomal or inter-chromosomal repeats including coding and non-coding tandem repeats, gene families and duplicated loci.

The focus of this section was the identification of variations in previously reported repetitive regions, such as the centromeres, and newly described in this study, such as transposable elements, and telomeric and subtelomeric regions.

3.2.4.1 Repetitive elements with homology to transposons

The analysis of repetitive regions in the genome of the natural isolate NRRL Y-11545, previously sequenced by Jeffries *et al.* (Jeffries *et al.* 2007), during this study revealed that repetitive elements with homology to Class I retrotransposons are found scattered along chromosome arms.

Our analysis has allowed the identification of a total of six different families, 3 novel *Copia* LTR-retrotransposons that have been named *Ava*, *Bea* and *Caia*, and 3 novel non-LTR LINE elements: *Ace, Bri* and *Can* (**Figure 3.6**).

All three elements identified contain only two ORFs: one encoding for POL and a *S. stipitis*-specific ORF that we named **L**TR-**A**ssociated **P**rotein (LAP1 associated with Ava, LAP2 associated with Bea and LAP3 associated with Caia) (**Figures 3.7, 3.8 and 3.9)**).

While Lap1 and Lap2 proteins are homologous, the Lap3 protein does not share homology with Lap1 or Lap2. Although in previous genome annotation reports Lap3 has been annotated as Histone H4 (Jeffries *et al.* 2007), our analysis demonstrates that Lap3

does not share any homology with others histone H4 proteins and does not contain a histone fold domain. Previous studies have already reported incorrect annotation of the *S. stipitis* genome (Maguire *et al.* 2013). Therefore, we conclude that Lap3 is not Histone H4 and that the gene has been erroneously annotated.

No GAG gene has been identified by homology analysis for *Ava, Bea* or *Caia* retrotransposons. Therefore, we hypothesise that the Lap proteins are non-canonical Gag proteins and that *Ava, Bea* and *Caia* are capable of transposition.

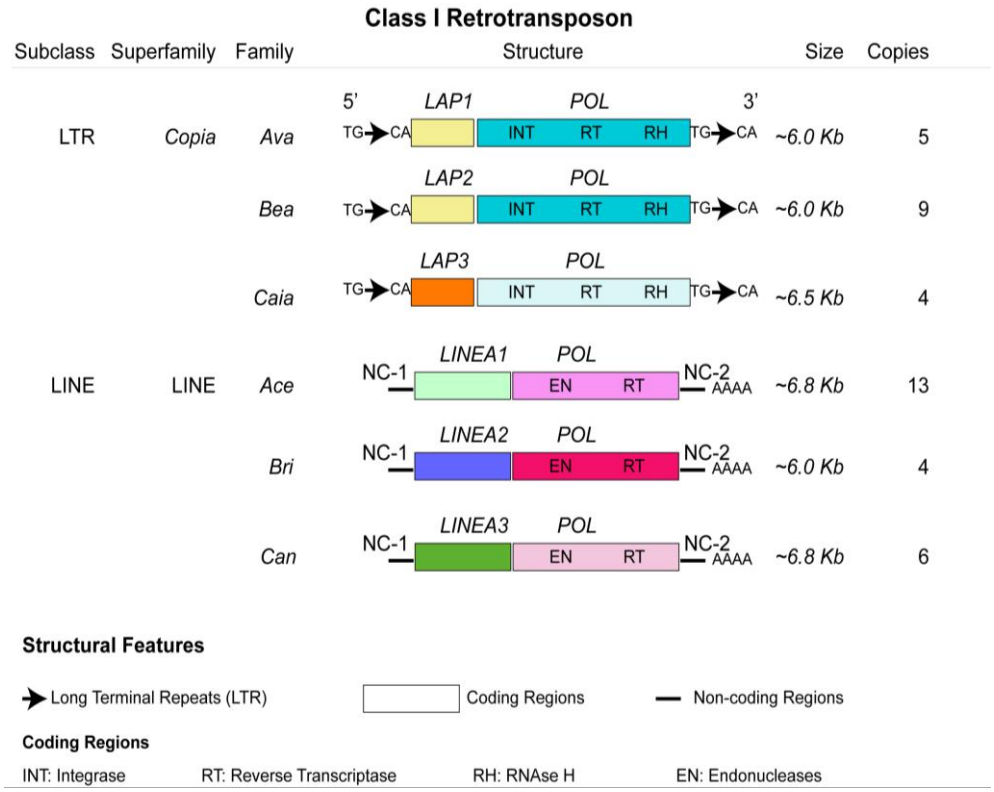


Figure 3.6. Classification of the families of class I retrotransposons identified in *S. stipitis* NRRL Y-11545.

Homology analysis were also used for the identification of 3 novel non-LTR LINE elements: *Ace, Bri* and *Can*. These elements share similar structure, since they are all surrounded by two Non-Coding regions: NC1 and NC2, with NC2 being associated with a terminal (3') poly(A) tail, which has been previously identified as a signature of LINE elements (Wicker *et al.* 2007). Moreover, the three elements also contain two ORFs: one of them codifying for POL, with both endonuclease and retrotranscriptase domains, whereas the other is a protein specific for each element (LINEA1 and LINEA2 and LINEA3 respectively).

Homology analyses identified the LINE-transposable element *Bri* as a *C. albicans* Zorro-3 like transposon. The Zorro-3 transposon in *C. albicans* is characterised by the presence of two ORFs and a terminal poly(A) tail in the 3'-UTR region. The first ORF (ORF1) presents two zinc fingers motifs, whereas the second ORF (ORF2) encodes for

an endonuclease and a reverse transcriptase. Moreover, it also presents a C-terminal zinc finger (Goodwin, Ormandy and Poulter 2001).

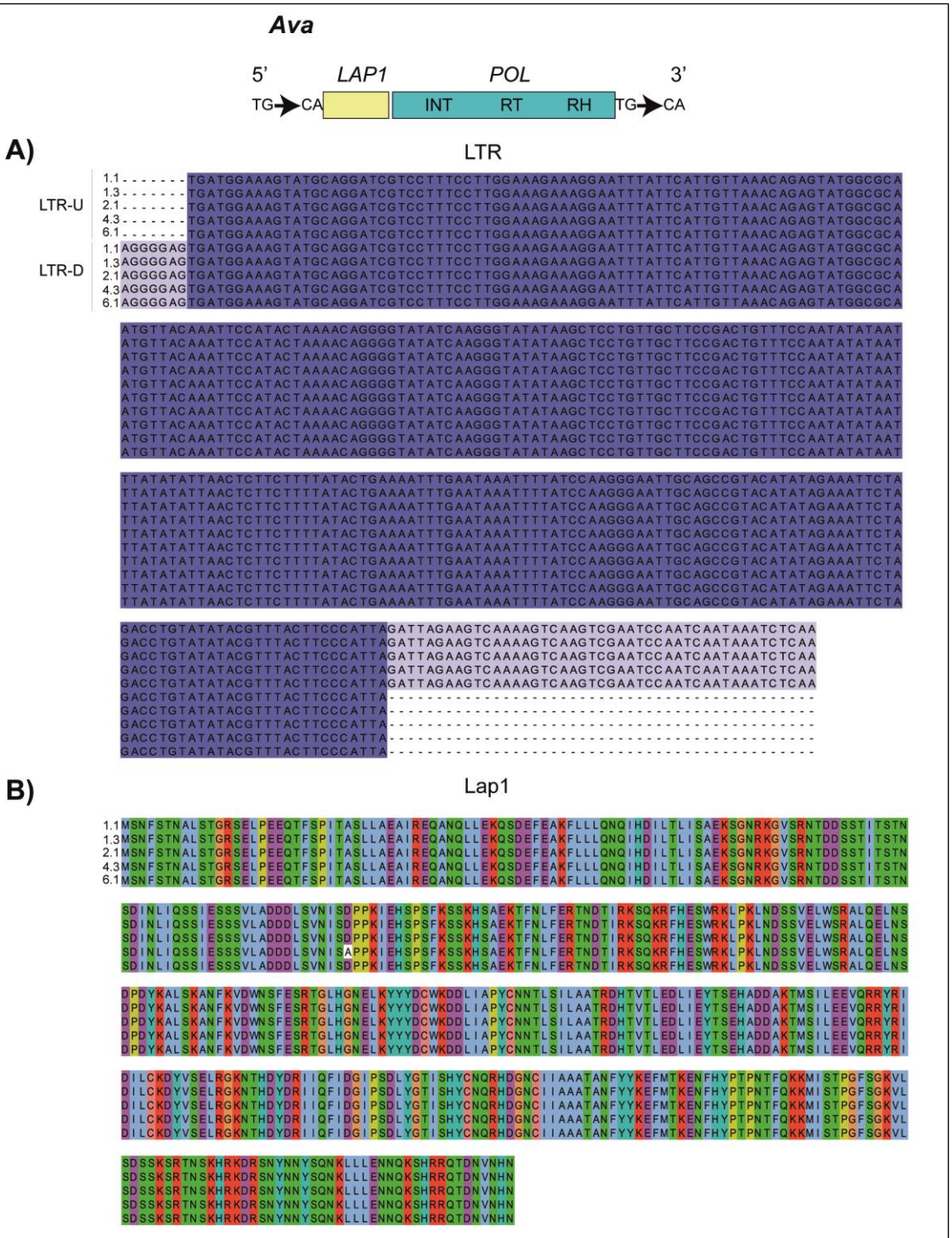


Figure 3.7. Alignment of elements associated to the TE LTR-*Ava* in the strain NRRL Y-11545. (A) Upstream (U) and downstream (D) long terminal repeats (B) Alignment of Lap1 protein.

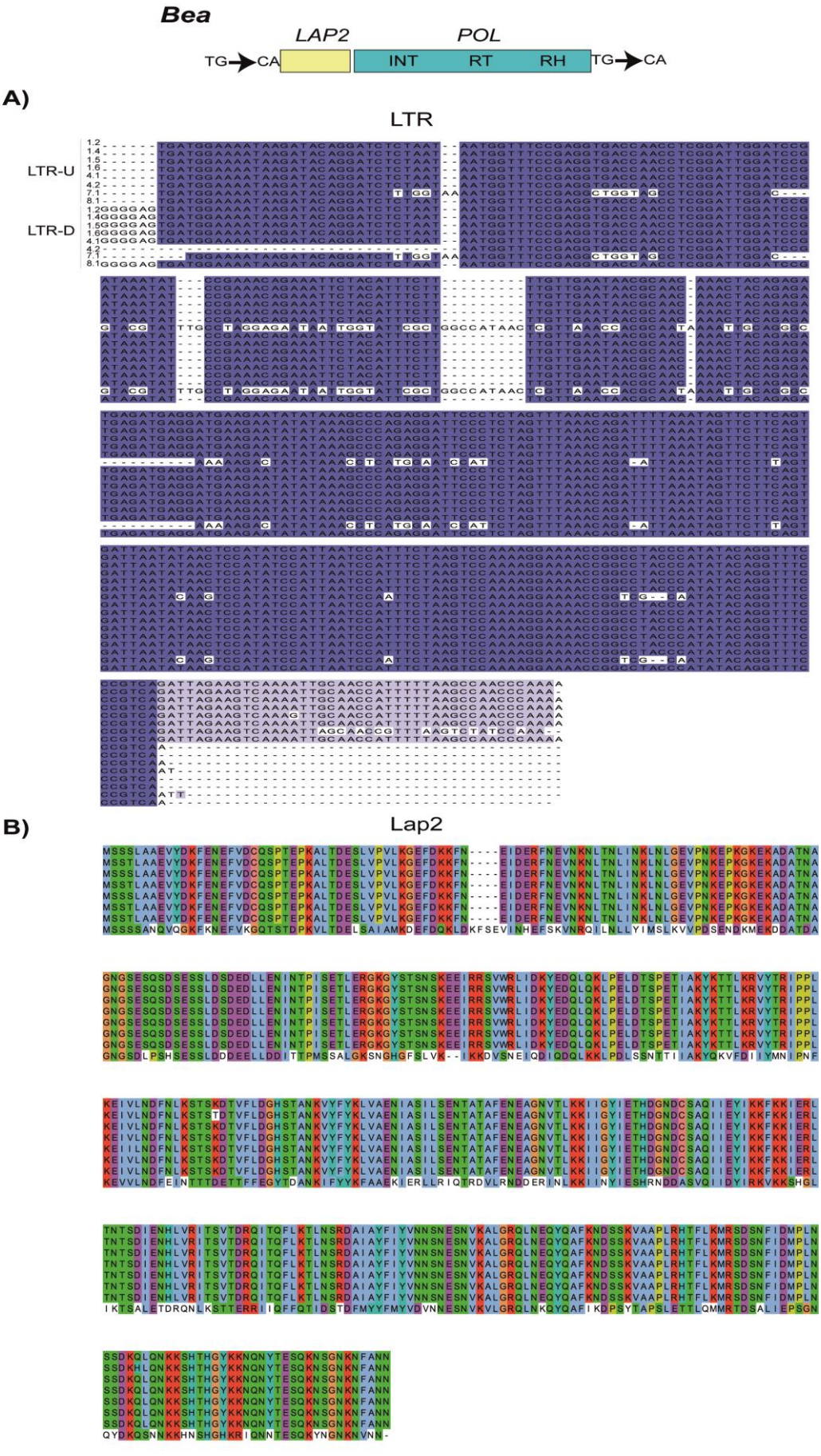


Figure 3.8. Alignment of elements associated to the TE LTR-*Bea* in the strain NRRL Y-11545. (A) Upstream (U) and downstream (D) long terminal repeats (B) Alignment of Lap2 protein.

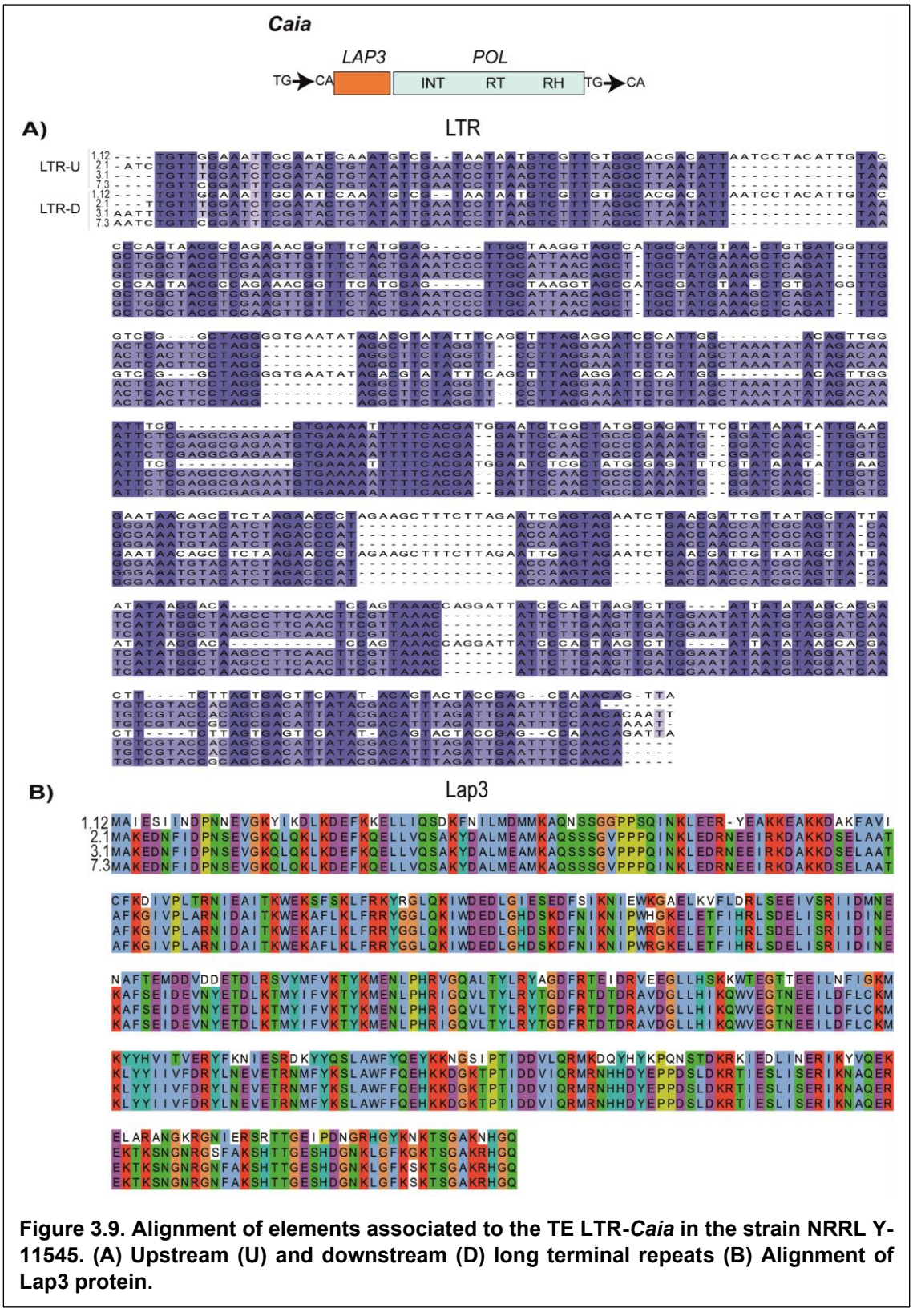


Figure 3.9. Alignment of elements associated to the TE LTR-*Caia* in the strain NRRL Y-11545. (A) Upstream (U) and downstream (D) long terminal repeats (B) Alignment of Lap3 protein.

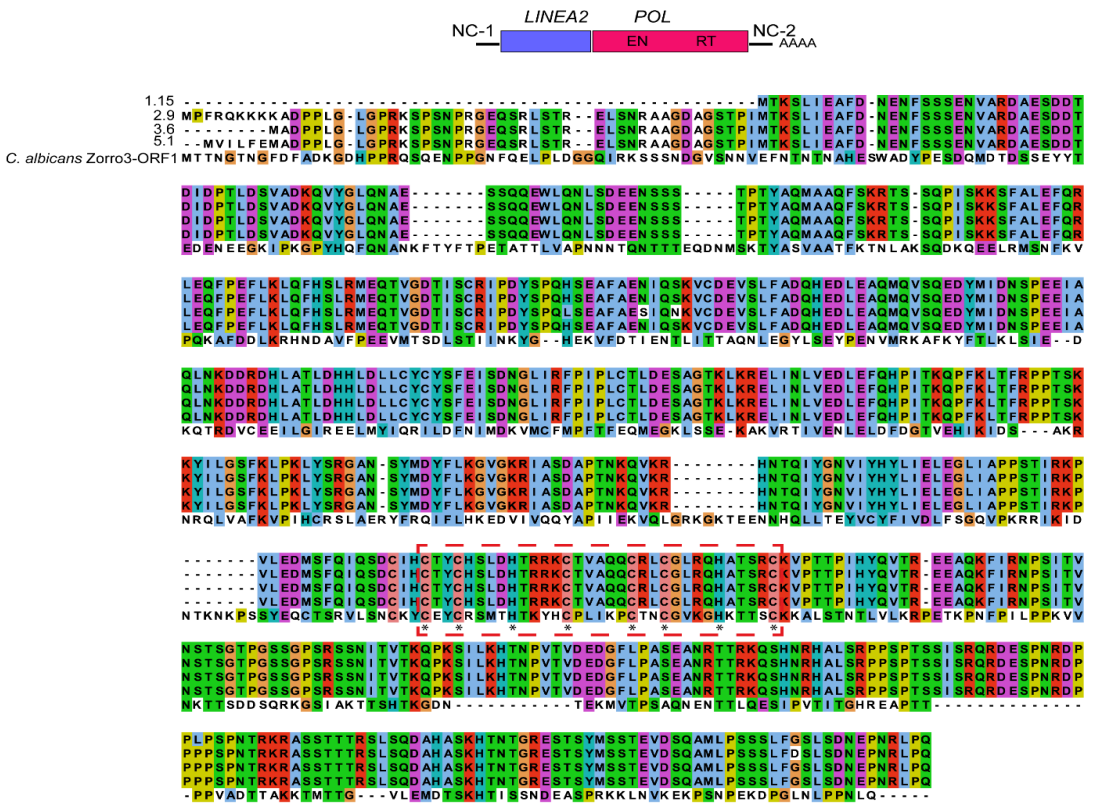


Figure 3.10. Alignment of LINEA2 ORF from *Bri* TE in *S. stipitis*. *C. albicans* ORF1 from Zorro3 transposon is added for comparison. (*) indicate residue conservation (C and H) in the zinc finger domains.

Alignments of the LINEA2 ORF show strong intraspecific sequence conservation in *S. stipitis* elements, although this conservation is lower when compared to *C. albicans* ORFs. Despite of this, the main domains described in *C. albicans*, two zinc fingers, are conserved in *S. stipitis* (**Figure 3.10**). Both endonuclease and reverse transcriptase domains are found in POL, however, the C-terminal zinc finger also described in *C. albicans* is not detected.

Two evidence led us to hypothesise that the remaining 19 TE can be further divided in two closely related groups, that to our knowledge had not been described to date.

First, BLAST analysis of the POL ORF present in all 19 TEs detects a ~30% identity (over 98% query covery) with the reverse transcriptase present in the budding yeast *Metschnikowia aff pulcherrima*. Although several strains have been sequenced and the gene models have been published (Venkatesh *et al.* 2018; Gore-Lloyd *et al.* 2019), to our knowledge, no studies on the transposon structure of *M. pulcherrima* have been conducted. Moreover, sequence alignments with the most common reverse transcriptase in TEs (Jockey, L1, R2 and RTE) fail to detect strong similarities (data not shown).

Second, alignment of the ORF1 proteins, associated to these elements shows two different consensus sequences (**Figure 3.11**). LINEA1, present in 13 TEs (1.8, 1.9, 1.10, 2.2, 2.7, 3.2, 3.4, 3.5, 4.6, 6.1, 8.1, 8.3, **Table S3**), has not been previously annotated, and is characterised by the presence of a coiled coil domain, a zinc finger, and a low complexity region. On the other hand, LINEA3 is present in 6 TEs (1.3, 2.8, 2.5, 4.1, 4.7 and 5.3, **Table S3**), and has been annotated as the gene *PSR2,* which codifies for a cytochrome c heme-binding site*.* However, protein homology comparison with *S. cerevisiae* Psr2 fails to detect sequence similarity. Moreover, the total size of LINEA3 is ~1000 aa, whereas Psr2 in *S. cerevisiae* is 347 aa. Finally, LINEA3 lacks the phosphatase domain present in Psr2. Therefore, we conclude that LINEA3 is not Psr2 and is also included in the genes incorrectly annotated (Maguire *et al.* 2013). No domains are detected in LINEA3, only 3 low complexity regions, so any prediction of its function is uncertain.

To unveil our hypothesis, a phylogenetic tree was constructed via a maximum likelihood model, using the reverse transcriptase domain of the POL ORF present in different LINE transposable elements. The domain sequences were aligned by Clustal-Wallis, and the maximum likelihood algorithm was used as test of phylogeny, with 100 bootstraps (**Figure 3.12**). The reverse transcriptase domain of the POL ORF detected in LTR transposable elements *Bea* previously described in this study was used as an outgroup.

The phylogenetic tree placed all non-LTR elements described in this study in the L1 clade. First, the transposable element *Bri* (green) forms a monophyletic group with *C. albicans* Zorro3. Furthermore, the uncharacterised 19 elements were classified in two different monophyletic groups, that we name *Ace* and *Can*. This classification also matches the differentiation according to the LINE specific proteins LINEA1 and LINEA3.

The distance observed between these elements in the tree indicate strong relationship. *Can* and *Ace* TEs present a 99% amino acid similarity over the POL ORF, although this similarity falls to 96% between *Can* and *Bri* and 88% between *Ace* and *Bri*, whereas L1s detected in humans and mice only share a 63% amino acid similarity.

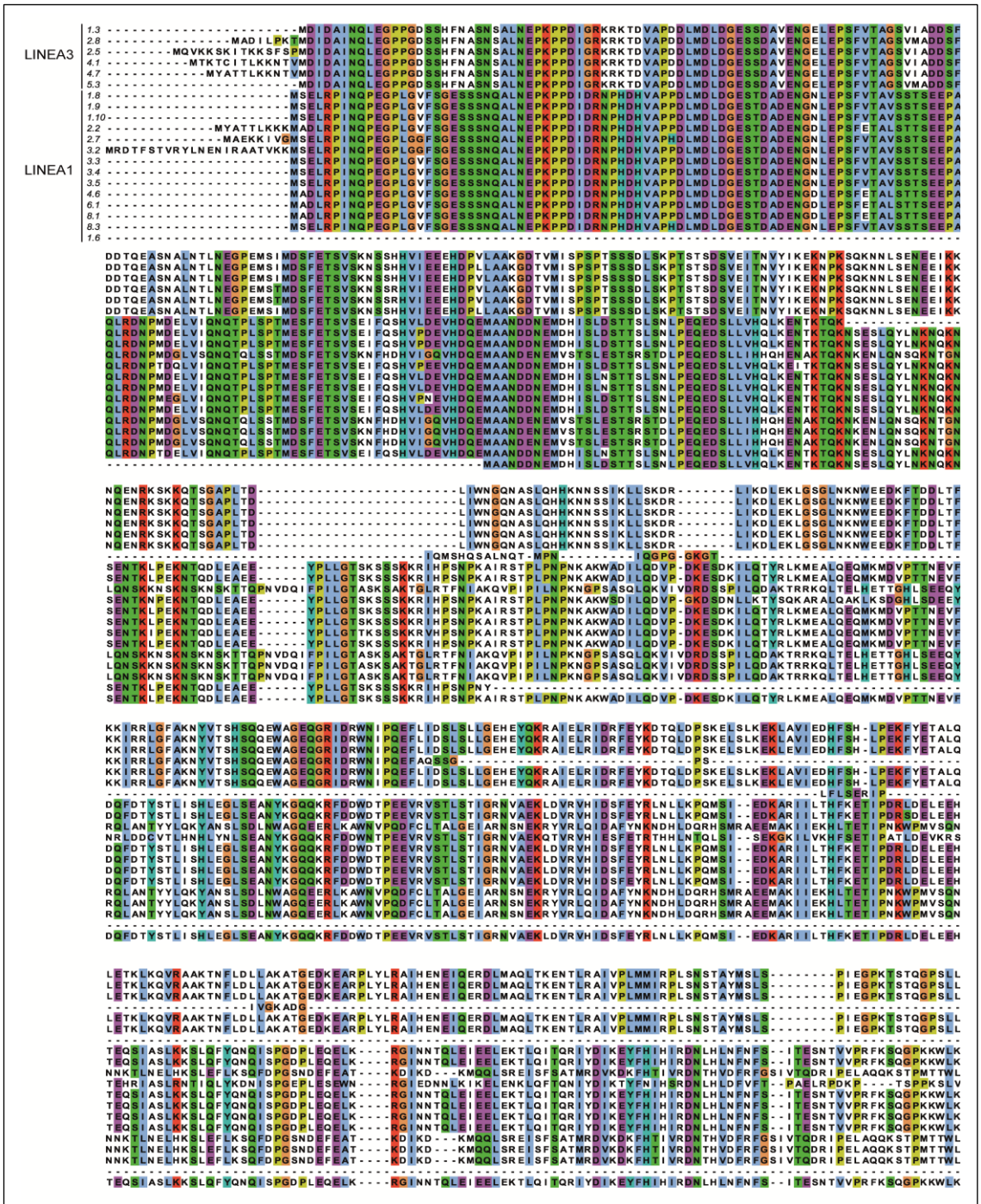


Figure 3.11. Partial alignment of LINEA proteins associated to the non-LTR transposable elements *Ace* and *Can*.

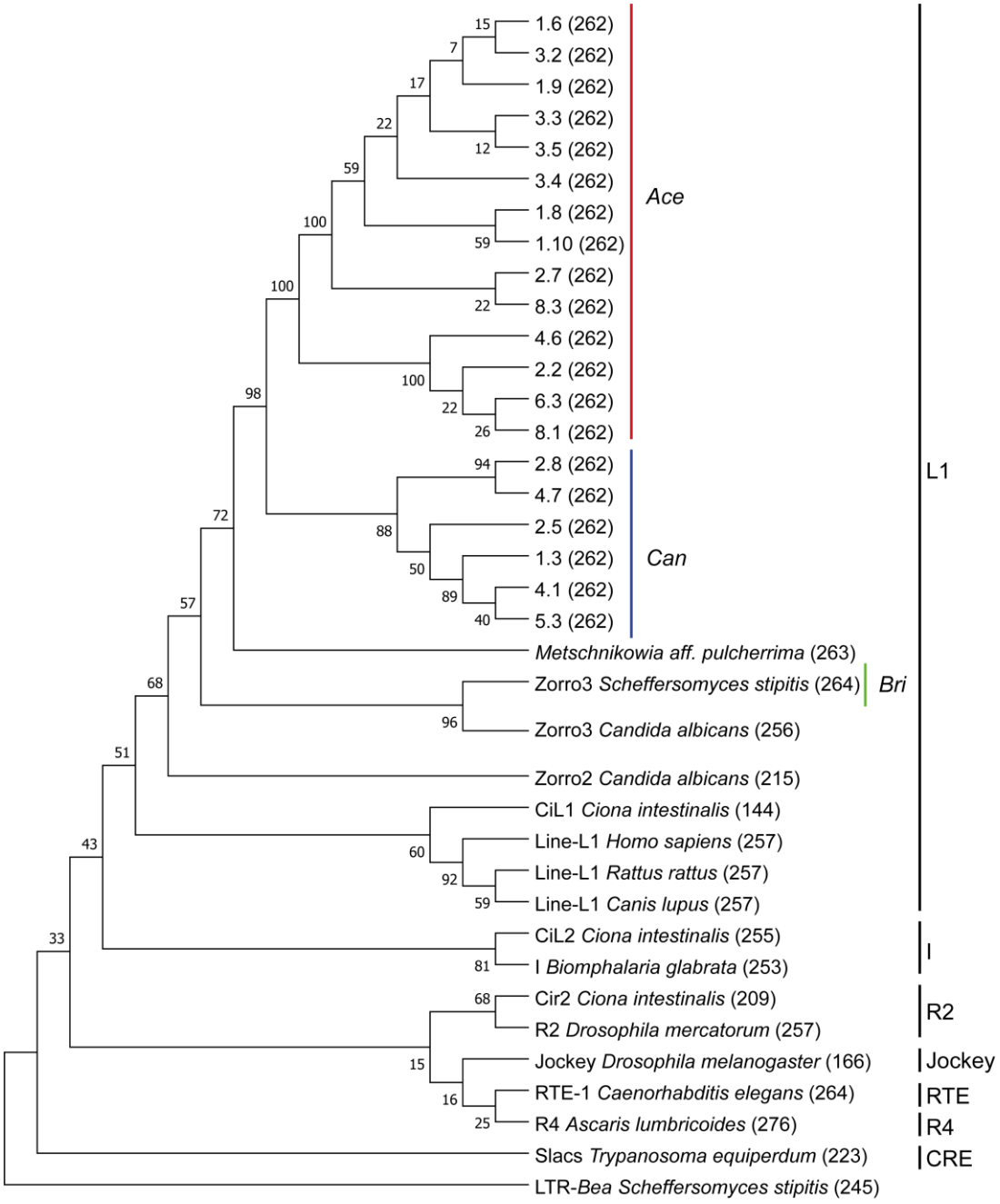


Figure 3.12. Maximum likelihood phylogenetic tree (100 Bootstraps) of the reverse transcriptase domain in the POL ORF of different LINE TEs. The reverse transcriptase domain of the LTR-*Bea* TE discovered in this study was used as an outgroup. The number in brackets indicates the length of the sequence (number of amino acids).

These data led us to conclude that a total of 6 novel transposons have been detected in *S. stipitis* natural isolate Y-11545 (3 LTR and 3 non-LTR (LINE, L1), of which two, *Ace and Can,* there is no previous description available to our knowledge.

The study of these TEs in the strain Y-7124 revealed that the 6 elements were present in both strains. However, one of the most prominent differences between the genomes of the two strains is in the abundance and localisation of these non-centromeric TEs and their transposons-associated repeats.

The number of non-centromeric retrotransposons and transposons-derived repeats is greater in the Y-11545 reference genome compared to the Y-7124 genome. Retrotransposons, both full length and truncated LTR and LINE elements account for approximately 2% of the reference Y-11545 genome and only for ~1% of the Y-7124 genome (**Figure 3.13, A**). Although the two natural isolates contain the same retrotransposons families, the Y-11545 genome contains 45 transposon copies, whereas only 20 transposons are associated with the Y-7124 genome. Most of these elements are found in different positions along the chromosome arms in the two isolates. Therefore, we classified retrotransposons loci present at the same position in both isolates (ancestral loci), present in the reference Y-11545 genome but absent in Y-7124 (deletion loci) and not present in the reference genome but present in a given strain (insertion loci) (**Figure 3.13, B**).

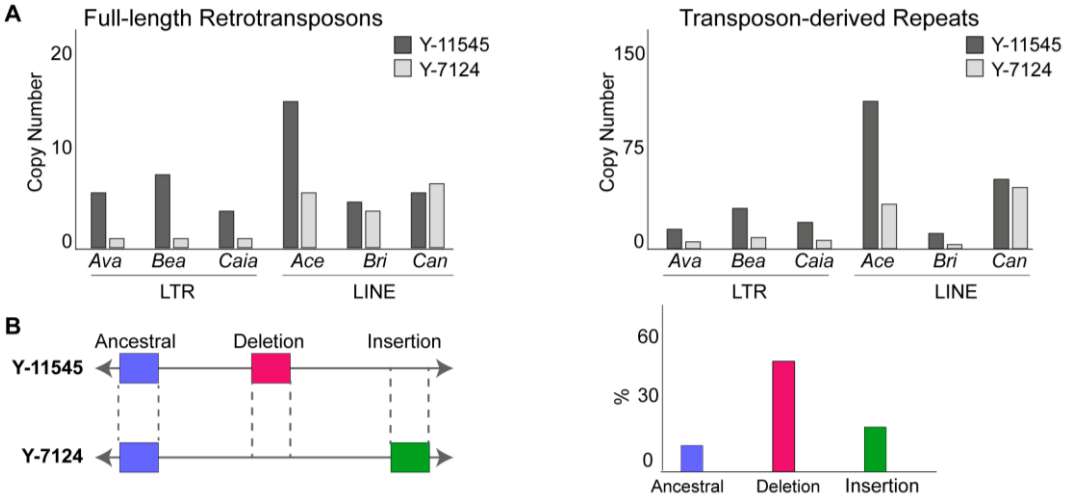


Figure 3.13. Differences in the TE distribution between two natural isolates of *S. stipitis* (NRRL Y-11545 and NRRL Y-7124). (A) Number of full length TEs and repeats associated to TEs per strain. (B) TEs conservation percentage. Ancentral TEs are present in both natural isolates, deletion TEs are only present in NRRL Y-11545 (reference) and insertion TEs are only present in NRRL Y-7124. Abbreviations: LINE: Long insterspaced nuclear element. LTR: Long terminal repeats.

Out of 69 transposons loci, only 10 ancestral loci (~15%) were detected in the two isolates. These sites, which comprise of LINE elements of the *Ace* family, are likely to be inactive transposons or transposons that rarely transpose. Sequence analyses of these reference loci confirm this hypothesis as 8/10 of the ancestral loci lack or contains a truncated POL ORF (data not shown). In addition, we detected 42 deletion loci (60 %) and 17 (24%) insertion loci (**Figure 3.13, B**). The presence of deletion and insertion loci suggests that *S. stipitis* LTR transposons and LINE elements are active and competent of transposition.

Active transposons can insert into genes to cause functional consequences. Comparison of transposon insertion between Y-11545 and Y-7124 reveals that no protein-coding genes were disrupted by a TE insertion. However, the presence of an *Ace*

LINE transposon is detected in the rDNA locus of NRRL Y-11545, disrupting the 5S rRNA from the rest of the locus. Remnants of the presence of a transposon are also found in NRRL Y-7124, but the full-length transposon is not detected (**Figure 3.14**).

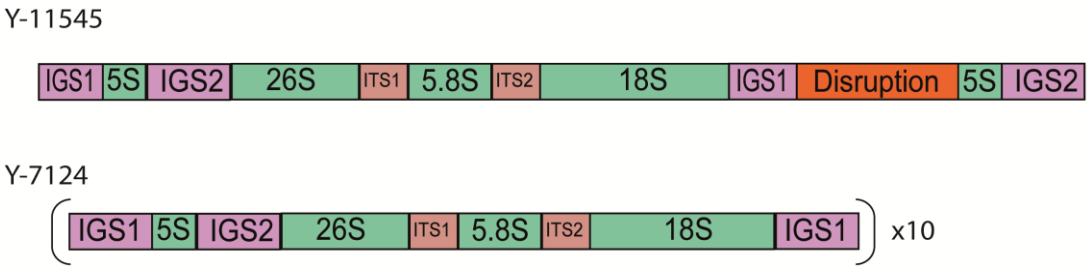


Figure 3.14. rDNA locus of *S. stipitis.* The presence of a TE is detected in the strain Y-11545, whereas only the remnants of a complete TE is detected in one of the copies differentiated in Y-7124.

In addition to the full-length elements, several (11) truncated versions of all the families are found across the genome of both strains (**Table S3 and S4**).

3.2.4.2 Genome diversity in *S. stipitis* centromeres

The structure of the centromeric regions of the *S. stipitis* reference strain Y-11545 has been previously studied by several groups. Both Jeffries *et al.* (Jeffries *et al.* 2007) and Lynch *et al.* (Lynch *et al.* 2010) proposed that centromeres are organised in G+C poor regions, which present clusters of Ty5-like LTR transposable elements (named Tps5). These elements can be classified in three different families (Tps5a-c), that present different terminal repeats associated (LTRa-c) (Coughlan and Wolfe 2019). Moreover, the presence of a repetitive 3.7 kb sequence, the large retrotransposition derivatives (or LARDs), formed by non-coding DNA and flanked by LTRs in parallel orientation (LTRd) has also been described.

Table 3.6. Summary of centromere size and number of centromeric elements present in *S. stipitis* strains NRRL Y-11545 and NRRL Y-7124.

	Y-11545 size (kb)	Y-7124 size (kb)		Y-11545	Y-7124		Y-11545	Y-7124
C1	36.4	31.8	Tps5a	10	2	LTRa	23	13
C2	36.0	18.6	Tps5b	14	3	LTRb	37	21
C3	41.0	16.2	Tps5c	4	0	LTRc	14	12
C4	29.2	22.4	LARD	6	4	LTRd	18	11
C5	22.7	21.4	Tps5ap	7	19	LTRat	15	16
C6	26.5	18.5	Tps5bp	24	41	LTRbt	11	7
C7	40.7	52.4	Tps5cp	3	7	LTRct	4	4
C8	43.6	22.2				LTRdt	2	3
average	34.5±7.6	25.5±11.8	Total	122	76	Total	124	84

The results of this study demonstrate that in both natural isolates the regional centromeres are distinct. First, the average size of the centromeres in the strain Y-7124 (25.5 ± 11.8 kb) was smaller compared to Y-11545 (34.5 Kb ± 7.6 kb) (**Table 3.6**).

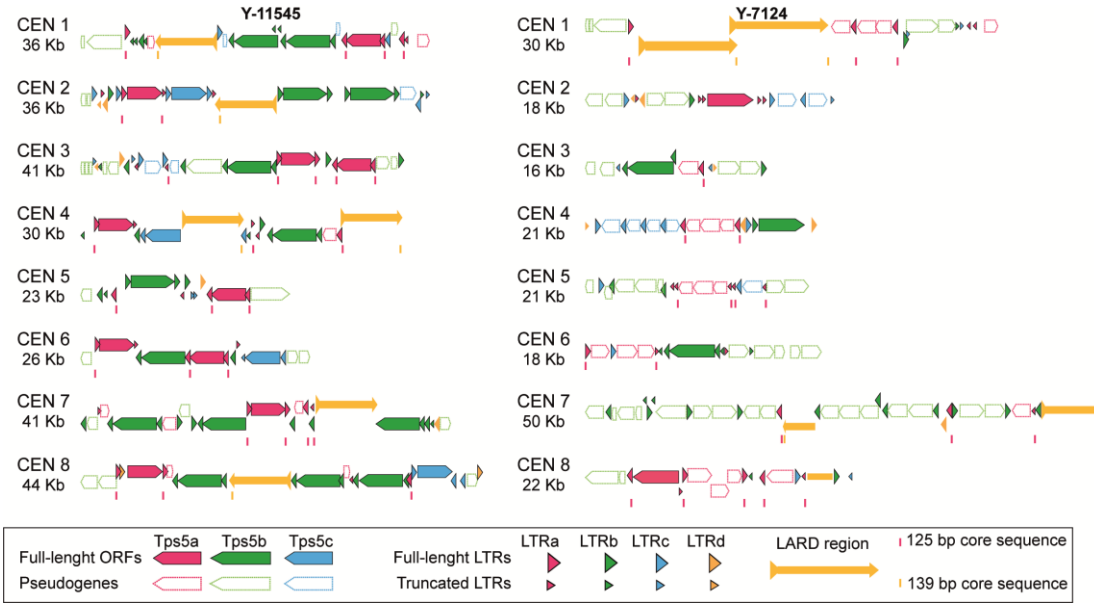


Figure 3.15. Centromere organization in *S. stipitis* natural isolates NRRL Y-11545 and NRRL Y-7124. Abbreviations: CEN: Centromere, LARD: Large retrotransposons derivatives, LTR: Long terminal repeat, Tps5: Ty-5 like retrotransposons.

Moreover, although the same family of Tps5 and LTRs are found in both species, their number and distribution in both genomes is distinctive (**Figure 3.15**). A total of 28 complete Tps5 genes were identified for Y-11545, whereas only 5 were identified in Y-7124 (**Table 3.6**). Both chromosome 5 and 7 of the later strain did not present any complete Tps5 ORF, which allows to conclude that their presence is not required for the formation of functional centromeres. Moreover, chromosome 2 in Y-7124 does not present any complete 125 bp core sequence, proposed by Cao et al (2017 and 2017) (Cao, Seetharam, *et al.* 2017; Cao, Gao, *et al.* 2017) as functional centromeres, which supports the rejection of point centromeres in *S. stipitis*.

The number of LARDs was also lower in the strain Y-7124, from 6 in Y-11545 to 4. Their presence is not detected in chromosomes 2, 4 and 8, whereas there is the presence of an extra LARD in both chromosome 1 and 5.

The final difference detected between the two strains is found in chromosome 7 of Y-7124, in which several Tps5b genes are found disrupting a LARD copy, separating the LTRd region that normally flanks it from the rest of the structure, event not observed in Y-11545.

3.2.4.3 *S. stipitis* telomeres and subtelomeres organisation

The terminal sequences of *S. stipitis* chromosomes are repeat-rich and composed of two elements with different degree of repetitiveness: telomere proximal sequences and subtelomeric regions.

S. stipitis telomeric sequences have not been described to date. Our analysis in both strains identified repetitive telomeric proximal regions that are different from canonical telomeres (composed of tandem arrays of 6bp GT motif) as they are formed by tandem repeats of a long 24 nt unit with unique sequence. Despite this sequence variation, *S. stipitis* telomeric repeats contain a T+G motif reminiscent of typical telomeric repeats (**Figure 3.16**).

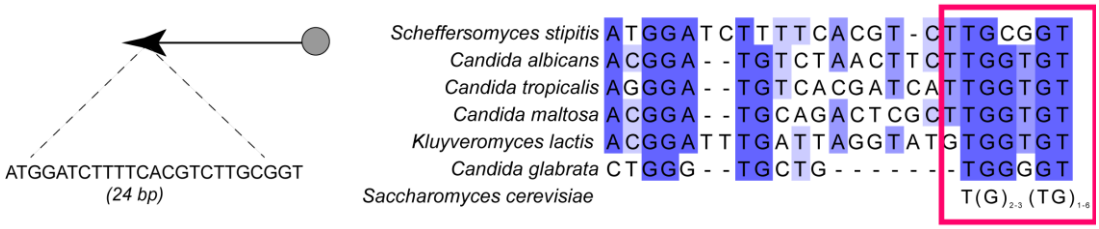


Figure 3.16. Repetitive chromosome-terminal sequences identified in *S. stipitis* NRRL Y- 11545. Alignment of *S. stipitis* telomeric repeats with other canonical telomeric repeats showing the presence of a conserved T+G motif.

S.stipitis subtelomeres (30 Kb regions adjacent to telomeric repeats) are enriched in retrotransposon-derived elements and gene families (**Figure 3.17**). Although no full-length retrotransposons are detected at subtelomeric regions, DNA sequences with homology to *Bea* LTR-retrotransposons (3) and *Ace* LINE elements (5) are found in 5/16 subtelomeres for the strain Y-11545 and to *Ace* LINE transposons (4) and to *Bea* LTR transposons (1) in 3/16 subtelomeric regions for Y-7124. Only the *Bea LTR* repeat was conserved in the C4 of both strains, although with differences in the repeat length (297 to 280 bp respectively).

S. stipitis subtelomeres are enriched in gene families, many of which are common subtelomeric genes (**Figure 3.17**) (Brown, Murray and Verstrepen 2010). The presence of genes encoding for predicted DEAD helicases (found in 7/16 subtelomeres), adhesins (10/16 subtelomeres), transporters of the MFS superfamily (8/16 subtelomeres) and genes encoding for predicted fungal specific transcription factors (9/16 subtelomeres) is detected for the strain Y-11545.

The number of subtelomeric regions in which DEAD helicases (13/16) and transcription factors (7/16) is different in NRRL Y-7124. Nevertheless, the amount of subtelomeric regions in which adhesins and transporters of the MFS superfamily is maintained (10/16 and 8/16 respectively) (**Figure 3.17**).

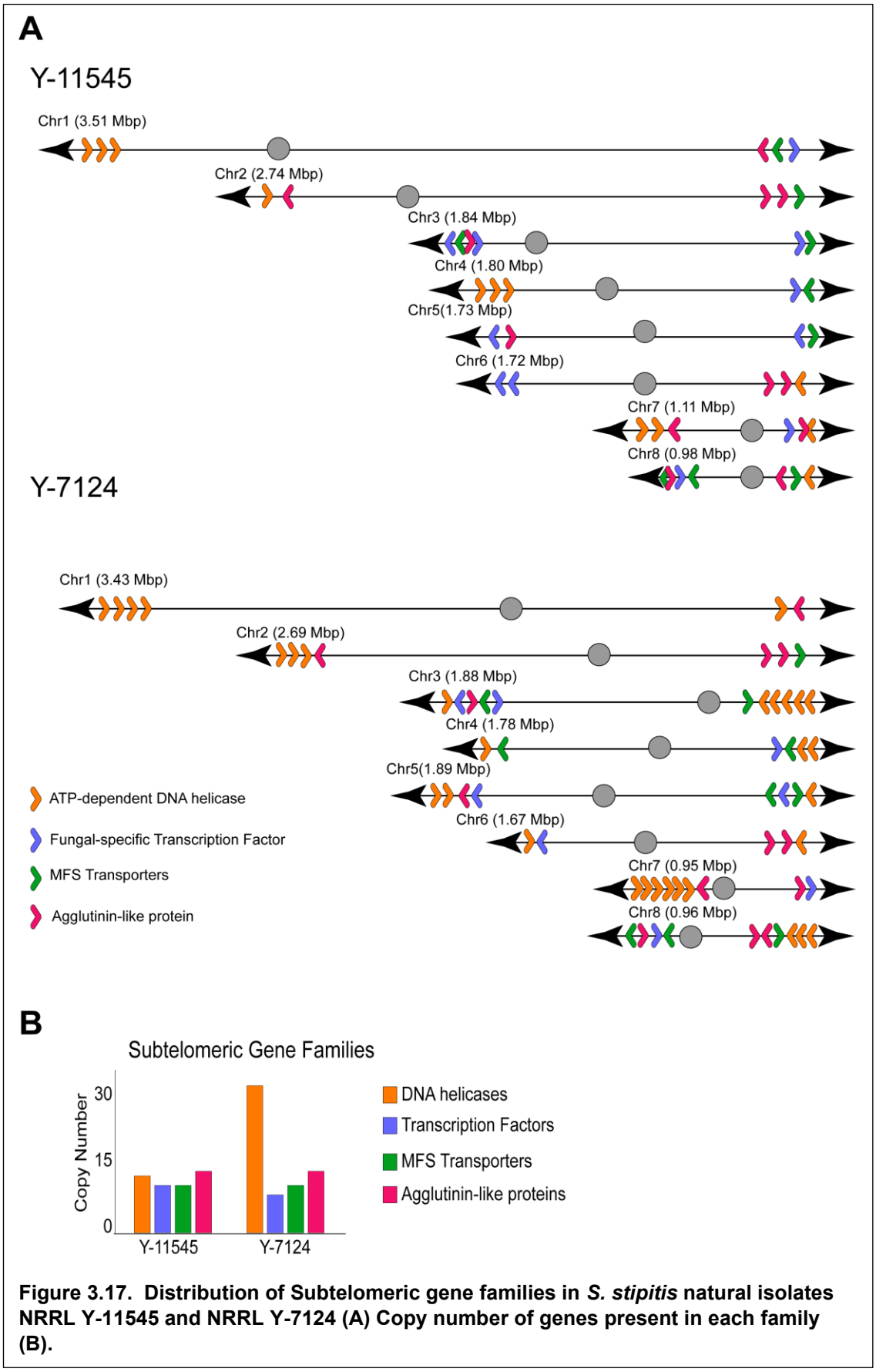


Figure 3.17. Distribution of Subtelomeric gene families in *S. stipitis* natural isolates NRRL Y-11545 and NRRL Y-7124 (A) Copy number of genes present in each family (B).

These differences between the two strains reveal that *S. stipitis* subtelomeric regions are plastic. Indeed, although overall the same gene families are found at subtelomeres, the number and position of genes families members varied between the two isolates with genes predicted to encode for ATP-dependent DNA helicases being the most variable (**Figure 3.17**).

3.2.5 Chromosome translocation between chromosome 5 and chromosome 7

Mummer alignments between the complete genome of the two natural isolates allowed the detection of one prominent genomic rearrangement: a reciprocal translocation between chromosome 5 and chromosome 7 (**Figure 3.18**). The translocation does not lead to truncation of any open reading frame and does not change genes orientation.

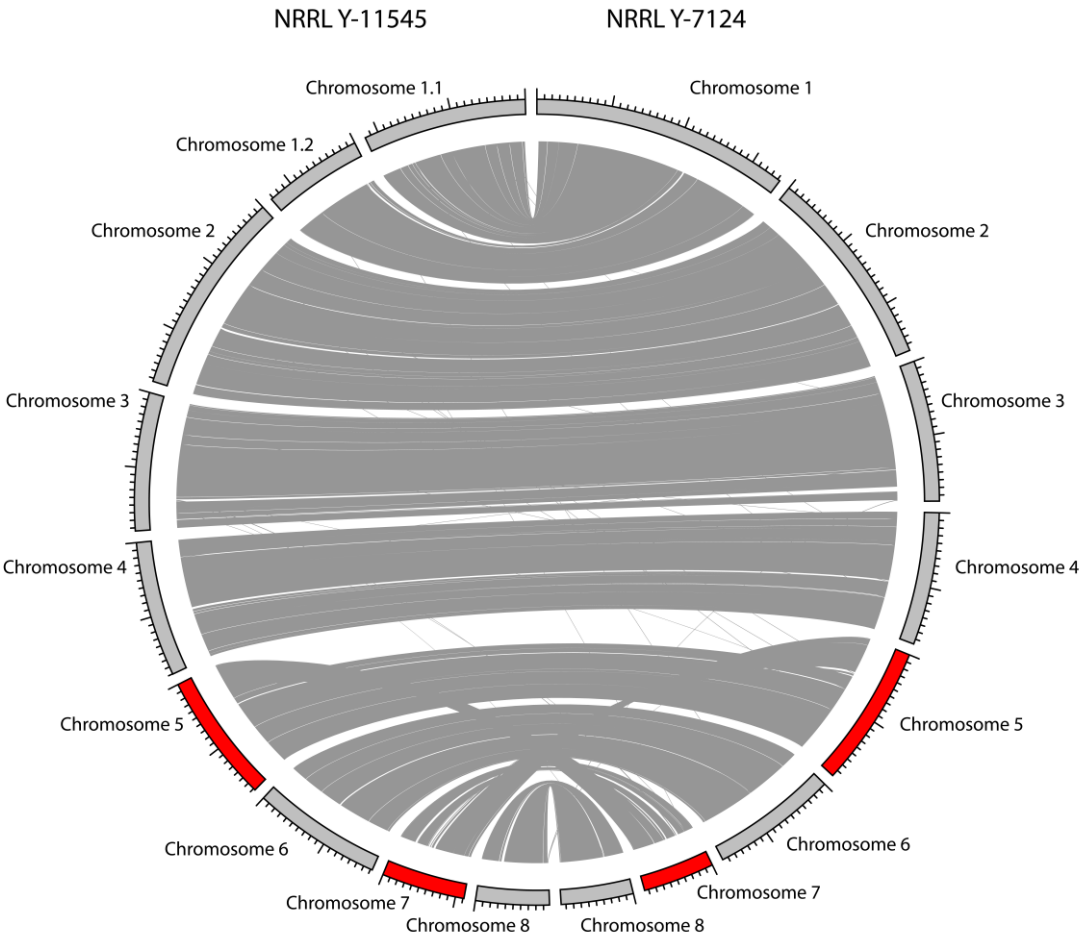


Figure 3.18. Mummer alignment between the whole genome of the *S. stipitis* strains NRRL Y-11545 and NRRL Y-7124.

However, a 17 KB region the chromosome 7 of the strain Y-11545 is not found in the genome of Y-7124. This region is mainly formed by inverted repeats (**Figure 3.19**), and only 6 ORFs are present (three duplicated): *PICST_63545* and *PICST_49558* (or *INPT2* and IPT1), which codify for inositol phosphotransferases, and also *PICST_33382* and *PICST_33387*, and *PICST_33383* and *PICST_33386* (*PSR1.2* and *PSR1.1*), which codify for 4 non-annotated proteins. These genes are present in the unique orthologs clusters identified for Y-11545 (**Table 3.4**) and are absent in the strain Y-7124.

As a result of this translocation 630 kb from the chromosome 7 of the strain Y-11545 are found on the right arm of the chromosome 5 of the isolate Y-7124. This translocation causes the size change in chromosome 5 and 7 of Y-7124 detected by CHEF karyotyping (**Figure 3.20**). Southern analyses with a probe specific for chromosome 5 of Y-11545 confirm this finding (**Figure 3.20**).

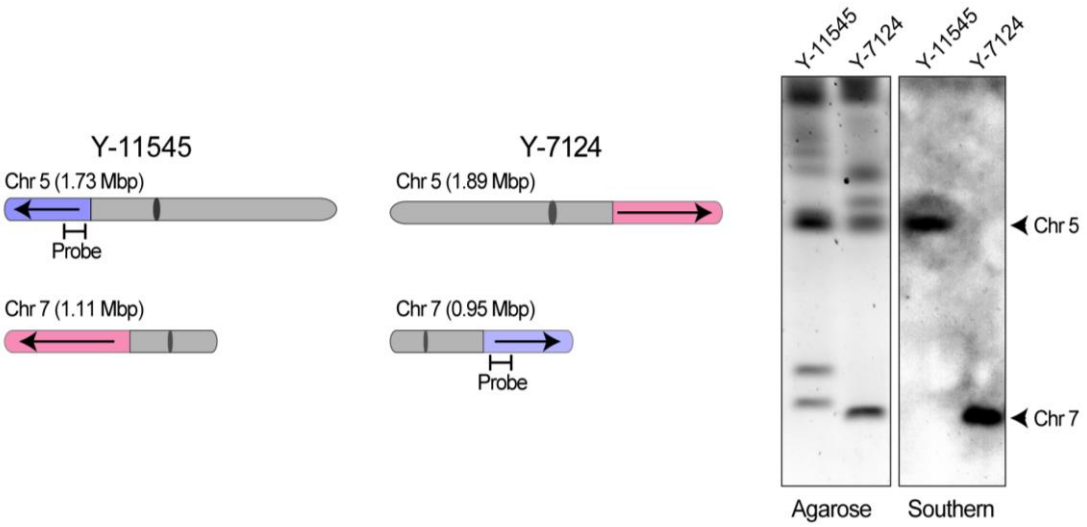


Figure 3.20. Southern Blot confirming the reciprocal translocation between chromosome 5 and 7 in the *S. stipitis* natural isolates NRRL Y-11545 and NRRL Y-7124.

The evolutionary history/relationship of Y-11545 and Y-7124 is unknown and therefore it is difficult to predict the exact molecular events and mechanism underlying this genomic change. However, sequence analysis of the rearrangement breakpoint reveals that this structural variation occurs in a genomic region that *(i)* contains homologous sequences between chromosome 5 and 7 and *(ii)* it is a transposon-rich region of chromosome 7 in Y-11545 (**Figure 3.19)**. The presence of transposons and transposon-derived repeats strongly suggest that these elements have mediated the chromosomal rearrangement. Therefore, we conclude that the translocation breakpoint is a structural variation hotspot that is marked by the presence of repetitive elements.

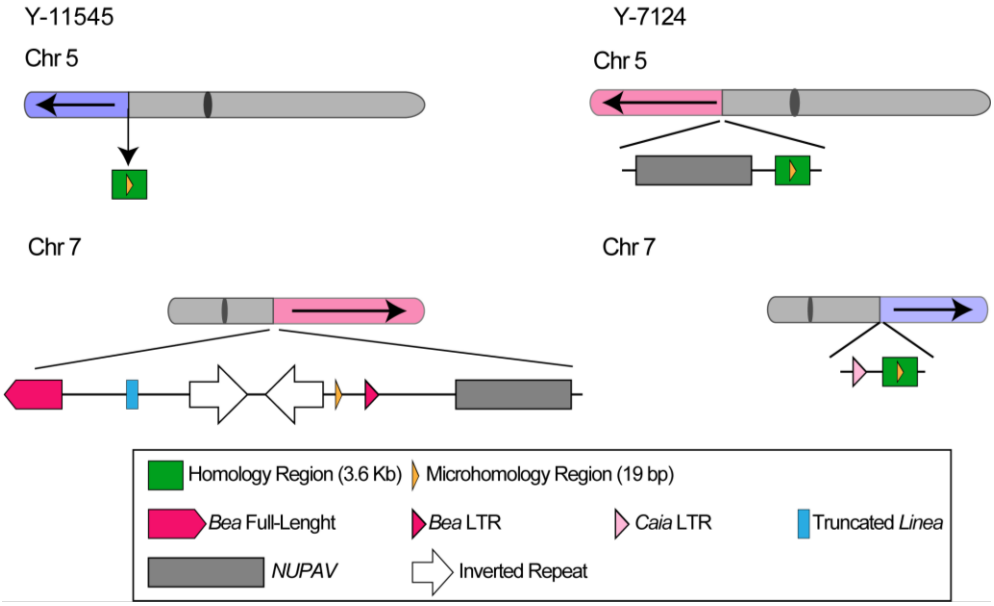


Figure 3.19. Instability hotspot and breakpoints involved in the reciprocal translocation between chromosome 5 and 7 in the natural isolates of *S. stipitis* Y-11545 and Y-7124.

3.2.6 Genome plasticity in laboratory adapted strains

The results of this study demonstrate that the genome of natural isolates of *S. stipitis* is characterised by wide plasticity. However, the evolutionary story of these strains is unknown, and therefore, the determination of how the genome has diverged results complicated.

To study the stability and plasticity of the genome, the strain Y-7124, commonly used in fermentation research, was adapted to temperature stress (37ºC) by streaking a colony growing in exponential phase into a new agar plate for 8 weeks. The media used for the experiment was SC, with a mixture of 60% glucose and 40% xylose as carbon source (SC-Mix), as it is the common sugar combination in lignocellulose medias. Growth at 30ºC was used as control.

Surprisingly, more genome modifications were observed in the strains grown as control (19/36), compared to the strains grown under temperature stress (12/34) (**Figure 3.21**). The most common modification was the presence of a novel band under 1.68 Mbp (25/31 strains presented this modification), which might indicate that this genome structure is the most stable among the new conformations.

These results suggest two possible theories: (i) The genome conformation of the natural isolate Y-7124 is not stable, and therefore the genome is quickly modified towards more stable structures, even when the conditions are not challenging (ii) The minimal media used as control (SC) might be more challenging than anticipated.

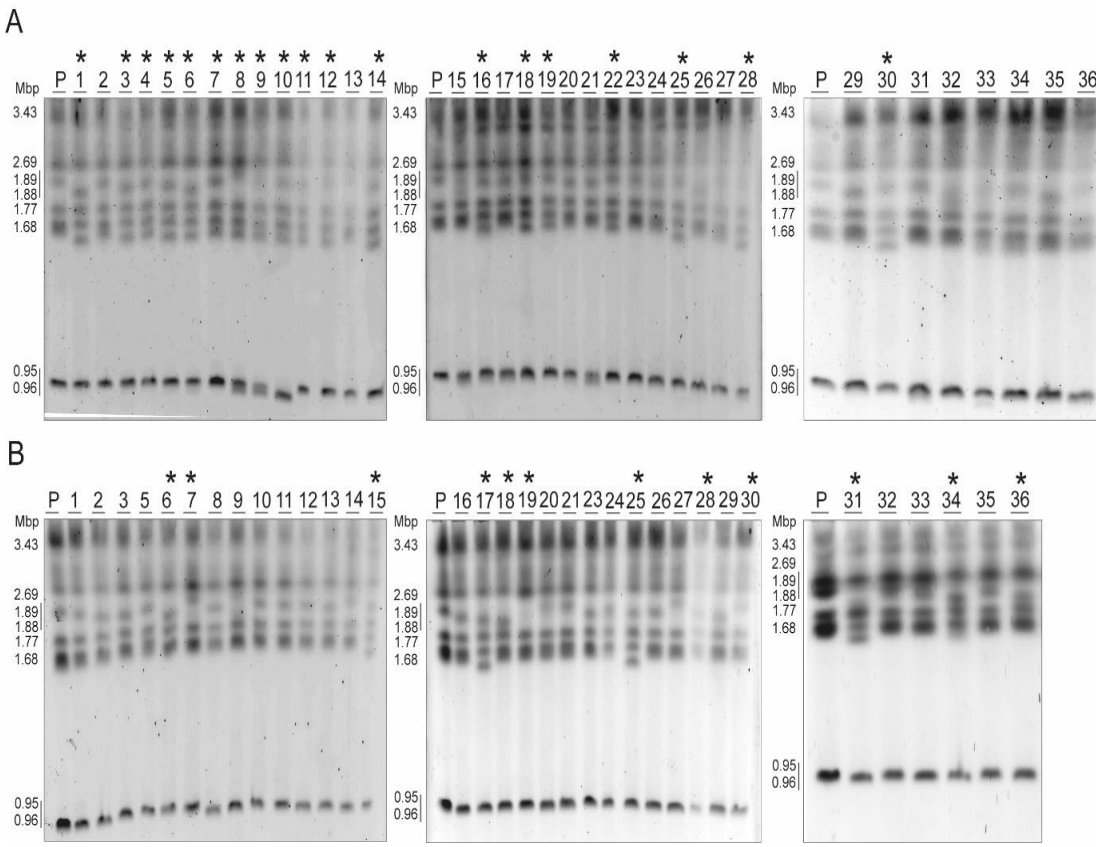


Figure 3.21. CHEF electrophoresis karyotype of the laboratory strains adapted from the natural isolate Y-7124 (P). (A) Control (B) Stress conditions (37ºC). (*) indicates isolates with karyotype modifications. Running conditions: 6 V/cm, switch time: 60-120s for 12 hours. 4.5 V/cm, switch time: 120-300s switch time for 12 hours at 4.5 V/cm. Staining in ethidium bromide (0.5 µg/ml).

Despite of the low resolution offered by CHEF electrophoresis; the wide genome plasticity observed in the evolution of natural isolates is also observed in laboratory adaptation experiments. Nevertheless, the nature of these modifications are not known, and therefore a deeper study on the genome structure of these strains is required.

3.3 DISCUSSION

Genome plasticity has been described as a characteristic of yeasts belonging in the CTG clade, such as *C. albicans* (Selmecki, Forche and Berman 2010; Legrand *et al.* 2019; Robert T. Todd *et al.* 2019). The aim of this chapter was to test and understand whether the genome of *S. stipitis*, a yeast present in the CTG clade, is also plastic. Previous studies conducted with 4 natural isolates of *S. stipitis* isolated from insect larvae (ATCC 58784, ATCC 58376, ATCC 62970, ATCC 62971) and 1 from a xylose fermenter (Y-17104) had observed small variations at chromosome organization by transverse-alternating field gel electrophoresis (TAFE) (Passoth *et al.* 1992). In that study, Passoth

et al. identified a total six chromosomes (bands) for all five strains, and the banding profiles could be classified in four different groups.

In this study, genome variability was assessed by a pulse field gel electrophoresis (PFGE) technique, CHEF electrophoresis. To our knowledge this is the first study that addresses the issue using strains from different isolation habitats and countries.

The results observed by Passoth *et al.* were reproduced successfully in this study. Moreover, the addition of 22 natural isolates allowed the identification of four new groups, which increased the number of different genome structures to a total of eight, indicating high genome plasticity in natural isolates of *S. stipitis*.

The first hypothesis of this study was that natural isolates that have been isolated from the same habitats might exhibit the same genome organization since their evolutionary pressure could be similar. To discover so, the relation between the two variables was studied with a chi-squared test (**Figure 3.22**), although, no relation was observed between the variables (p-value=0.791).

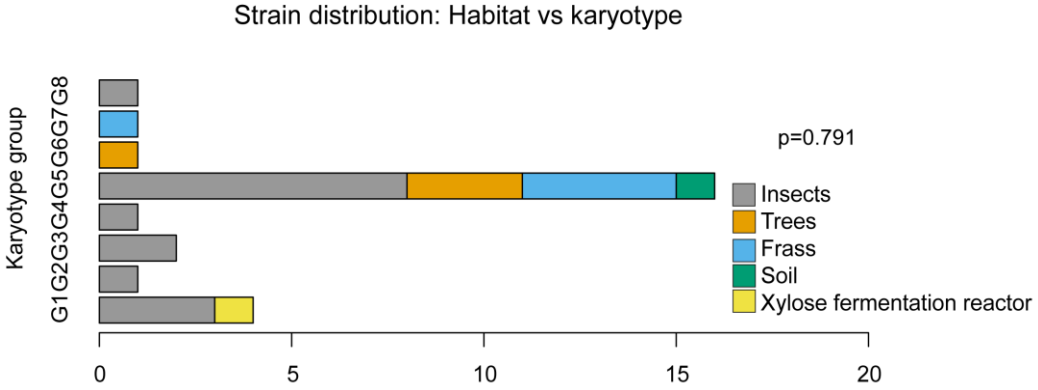


Figure 3.22. Distribution of *S. stipitis* natural isolates according to their karyotype (CHEF group, G1-G8) and their habitat: insect (grey), trees (orange), frass (blue), soil (green) or xylose fermentation reactor (yellow). No relation between the two variables was observed (Chi-squared test, p<0.05).

Despite the lack of connection between the genome organization and the habitat from which the strains were isolated, wide genome plasticity is observed. However, the reason why *S. stipitis* exhibits plasticity is not known.

Commonly, yeast species from the CTG clade that exhibit these levels of modifications are characterised by the presence of repetitive regions in their genome, which will lead to instability and genome shuffle under stressful conditions (Freire-Benéitez, Price, *et al.* 2016; Robert T. Todd *et al.* 2019; Dunn and Anderson 2019). To further investigate this, the repetitive regions of the only strain of *S. stipitis* sequenced and assembled to chromosome level to date, NRRL Y-11545, were characterised. Moreover, another natural isolate with different genome organization (NRRL Y-7124, a

strain commonly used in *S. stipitis* fermentation research) was sequenced by a hybrid TGS approach, using both ONT and Illumina, and its repeats were also characterised and compared to the strain Y-11545.

Three main elements were identified as repetitive DNA providers in the genome of *S. stipitis*: the centromeres, subtelomeres and transposable elements.

Transposable elements are the major contributors to repetitions in *S. stipitis* genome. The contribution of transposons to high number of repetitive regions has been reported previously for many organisms (Bleykasten-Grosshans and Neuvéglise 2011). However, to our knowledge, they had never been characterised in *S. stipitis*. This study identified 6 novel transposable elements (3 of them LTR and 3 non-LTR) as part of the genome of the natural isolate Y-11545.

LTR elements *Ava, Bea* and *Caia* share a common structure, characterised by the presence of long terminal repeats marked by 5' TG and 3' CA dinucleotides, which flank two ORFs. This structure has been previously described in the LTR transposable elements of other organisms and which the two ORFs consist of a gag and a pol protein (TyA and TyB in *S. cerevisiae*) (Bleykasten-Grosshans, Friedrich and Schacherer 2013). Despite the maintenance of the overall structure, no GAG gene was detected for any of the LTR elements in *S. stipitis*. Instead, 3 specific proteins were identified as part of the *Ava, Bea* and *Caia* elements (LAP1, LAP2 and LAP3 respectively). No specific domains are found in any of the proteins, so the prediction of their activity is uncertain. Nevertheless, the existence of strong variability between gag (TyA) proteins in different *S. cerevisiae* strains has been previously reported (Jordan and McDonald 1998; Kim *et al.* 1998; Bleykasten-Grosshans, Friedrich and Schacherer 2013). Therefore, their constant presence in the elements and lack of similarity with other gag proteins led us to hypothesise that LAPs are non-canonical GAGs.

This study also identified 3 novel non-LTR transposons in *S. stipitis*, *Ace, Bri* and *Can,* all of them belonging to the family L1, inside the superfamily LINE.

Bri transposons present a 96% of amino acid sequence similarity to the Zorro3 reverse transcriptase, only described previously in *C. albicans* (Goodwin, Ormandy and Poulter 2001). Goodwin *et al.* reported that the low sequence similarity between the 3 different Zorro transposons they found in *C. albicans* (less than 30%, compared to over 60% similarity between L1 reverse transcriptase of humans and mice) might be related to a different evolutive history, which suggest that their presence in the genome is ancient. Our phylogenetic analysis supports this theory, since *Bri* elements in *S. stipitis* and Zorro3 in *C. albicans* are homologous, which suggest their presence in a common

ancestor before the whole genome duplication event in yeasts that differentiates the evolutionary history between *C. albicans* and *S. stipitis*.

Ace and *Can* transposons also belong to the non-LTR transposon LINE family L1. They are characterised by differences in both ORFs present in the element, although their evolutionary divergence is recent compared to *Bri* elements. Remarkably, to our knowledge, *Ace* and *Can* transposons have been described for the first time in yeasts in this study, and their presence in other yeasts (such as *M. pulcherrima*) is also suggested.

After the detection and classification of transposable elements in the strain Y-11545, their identification was also carried out in the other *S. stipitis* natural isolate under study, Y-7124. The same families of transposons were identified in the two strains, although both the total number of full-length transposable elements and repetitive regions associated to them were higher in the strain Y-11545 (with a total of a 45 complete TEs and repetitions that account for a 2% of the genome in Y-11545 and 20 complete TEs and repetitions that account for approximately 1% of the genome in Y-7124).

Only 10 loci were present in both strains (~15%), which suggest that they are elements that rarely transpose. This is supported by the presence of truncated POL ORFs in these elements. The rest of the transposons (~85%) are present in only one of the strains, which suggests that they are still active and competent of transposition. Northern Blot experiments have been previously used to demonstrate the expression and activity of transposons (Goodwin, Ormandy and Poulter 2001; Deininger and Belancio 2016), although this study does not contain further experiments regarding transposons activity.

The importance of identification of transposons and repetitive regions in the genomes lies on the effects they might produce.

First, transposons can modify the gene expression upon integration. When a transposon is integrated in next to an ORF in the opposing direction, its promoter will control the expression of the ORF, and therefore it will depend on the signalling pathways that promote transposon expression. Examples of this have been previously reported in *S. cerevisiae*. For instance, adenine deprivation activates Ty1 transcription and retroposition (Todeschini *et al.* 2005), and their integration has been proven to condition the expression of genes adjacent to the insertion point under this severe adenine deprivation (Servant, Pennetier and Lesage 2008). Our study has identified only one gene disrupted by a transposable element, whereas the rest seem to be integrated in non-coding DNA. The disruption is observed in one copy of the rDNA gene by an *Ace*

TE. This is identified in both strains Y-11545 and Y-7124, although the element is more degenerated in Y-7124, indicating more mutation rates in the area for the later strain.

Second, transposons and other repetitive elements are associated with gross chromosomal rearrangements, including deletions, segmental duplications, inversions and reciprocal and non-reciprocal translocations (M. J. Dunham *et al.* 2002; Umezu *et al.* 2002; Chan and Kolodner 2011). This seems to be related with DSB repair (Argueso *et al.* 2008) linked to the formation of non-canonical DNA structures, such as hairpins or R-loops (Casper *et al.* 2009), and also with homologous recombination (Argueso *et al.* 2008).

Centromeres are structures whose complexity has been widely reported, from the point centromeres (~125 bp) described in several budding yeasts, such as *S. cerevisiae,* to epigenetically defined centromeres, present in most other organisms, that can span from few kilobases (like in *S. pombe,* or *C. albicans*) to hundreds of kilobases in most plants and animals (Malik and Henikoff 2009).

The centromeres of *S. stipitis* have previously been reported as epigenetically defined, formed by repetitive DNA, low G+C content, and rich in full length LTR-transposons and transposons derivative genes and repetitions (Coughlan and Wolfe 2019). Centromeres rich in transposons have also been previously described in other organisms, such as maize (Topp, Zhong and Dawe 2004) or marsupials (Carone *et al.* 2009).

Despite of the initial controversy created by the publications by Cao *et al*; (Cao, Gao, *et al.* 2017) (Cao, Seetharam, *et al.* 2017)) reporting a 125 bp sequence as point centromeres in *S. stipitis*, Coughlan and Wolfe (Coughlan and Wolfe 2019) demonstrated that this sequence is actually part of one of the LTR associated to a transposon and corroborated that *S. stipitis* centromeres are regional. The results presented in this study confirm this finding, and also rules out the compulsory requirement of a complete 125 bp sequence to constitute an active centromere, since none is detected in the centromere of chromosome 2 of the natural isolate Y-7124.

Moreover, this study has demonstrated that the size and organization of the centromeres in these two natural isolates of *S. stipitis* is distinctive, with larger and more transposon rich centromeres in the strain Y-11545. This indicates the action of evolutionary forces in the shape of centromeres in *S. stipitis.* These findings agree with the observation of rapid evolution in point and epigenetically defined centromeres of other yeasts, such as *C. albicans* (Padmanabhan *et al.* 2008) or *S. cerevisiae* (Bensasson *et al.* 2008), or even in humans (Rudd 2005), by the action of both recombination and sequence mutations.

Finally, chromosome ends are also known as repetitive-rich regions in the genome. The presence of the telomere repeats proportions chromosome stability. Moreover, subtelomeric regions (<30 kb distance from telomeric repeats) have also been reported as repetitive regions in the genome of several organisms (Sadhu *et al.* 1991; Louis 1995; Vershinin, Schwarzacher and Heslop-Harrison 1995; Torres *et al.* 2011). Our study detected multiple copies of genes of different families, commonly described as subtelomeric genes: DEAD helicases, major facilitator superfamily of transporters and fungal transcription factors (Brown, Murray and Verstrepen 2010), and differences in the number and orientations of these genes were also detected between the two natural isolates of *S. stipitis*, which suggests plasticity in the subtelomeric regions. Extensive interchromosomal reshuffling has been reported for different strains in other yeasts, such as *Saccharomyces paradoxus* (Yue *et al.* 2017) or *S. cerevisiae* (Yue *et al.* 2017), where subtelomeric rearrangements depend on recombination processes that promote joining of both homologous and heterologous sequences (Ricchetti, Dujon and Fairhead 2003). These effects have also been described in bacteria (Thibessard and Leblond 2014) and even in humans (Flint *et al.* 1995).

Despite this, the changes in the number of genes present in the subtelomeric regions of both strains can be controversial. First because the assembly of the strain Y-7124 conducted in this study split several ORFs in the subtelomeres, generating multiple ORFs that in Y-11545 are assembled as one, and therefore, leading to an erroneous number count. Moreover, the genome of the reference strain Y-11545 was sequence by a shotgun approach, based only in short DNA fragments. It has been reported that genome assemblers have problems to differentiate the correct positioning of repetitive reads (Tørresen *et al.* 2019). Therefore, the possibility that the assembly of the reference strain Y-11545 has placed all the reads for helicases in the same 7 subtelomeres (instead of 13 in Y-7124) cannot be ruled out. This could be easily checked with a PCR approach designed to detect helicases in all 16 subtelomeres.

Repetitive regions have been previously associated with chromosome rearrangements. Our study has identified one reciprocal translocation between the chromosome 5 and 7 in the natural isolates of *S. stipitis* NRRL Y-11545 and NRRL Y-7124. This translocation seems to be related with repetitive regions in the chromosome 7 of the strain Y-11545. The region rich in repetitions (both transposon and non-transposon related) accounts for a total of 136 kb (12% of the total chromosome length). The existence of transposons hotspots has been previously described in larger genomes, such as maize (SanMiguel *et al.* 1996). Additionally, this region presents a 17 kb fragment formed by inverted repeats in the strain Y-11545 but it is missing in the strain Y-7124. Moreover, 5 genes from viral origin (4 gag proteins and 1 RNA polymerase),

previously described by Frank and Wolfe (Frank and Wolfe 2009) are also present. Frank and Wolfe named these elements NUPAVs (Nuclear sequences of plasmid and viral origin), and they hypothesised that they were incorporated during the reparation of a double strand break in several yeast species. The identification of a chromosome translocation nearby in this study enforces this theory. Nevertheless, the NUPAV incorporation could have been prior to the translocation identified, since it is present in the chromosome 7 of the strain Y-11545, but in the chromosome 5 of Y-7124.

This evidence led us to hypothesise that the divergence of these strains occurred due to the instability caused by the repetitive areas in the chromosome 7 of a precursor strain (**Figure 3.23**). However, to confirm this, more evidence is required, and the whole genome sequencing of other natural isolates would help to elucidate the existence of strains whose genome structures could be intermediate in the evolutionary history presented in our model.

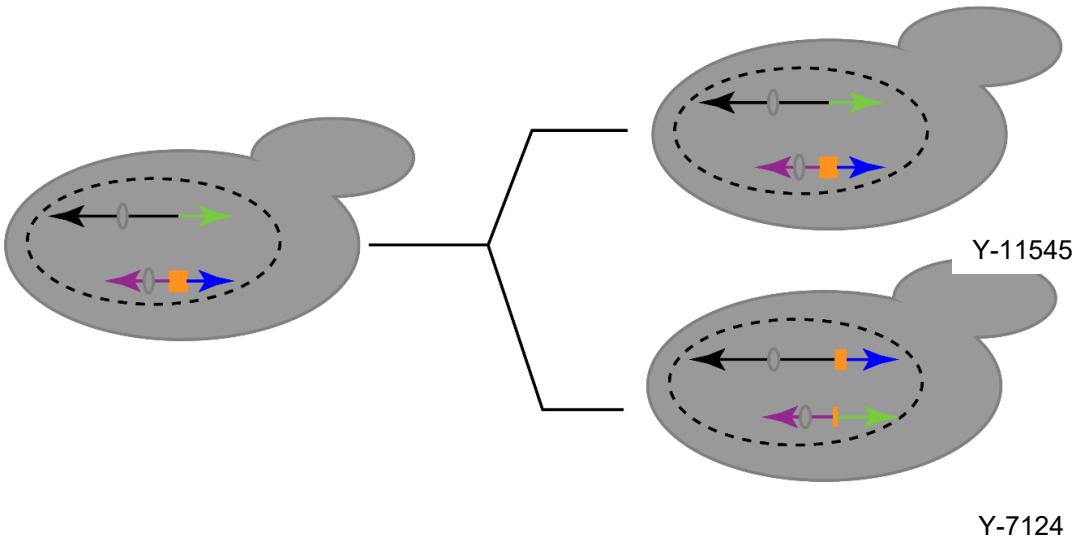


Figure 3.23. Evolutive model in which the instability associated to the chromosome 7 in a common ancestor caused a reciprocal translocation that originated the evolutive divergence between the *S. stipitis* strains Y-11545 and Y-7124.

Laboratory scale evolution experiments can easily be carried out and used as a model of modifications that might be observed in real-time evolutive history. Studies on karyotype variations of strains that have been grown under selective pressures have been conducted extensively in brewing industry (Deželak *et al.* 2014; Mangado *et al.* 2018; Large *et al.* 2020), in clinical research (Bravo Ruiz *et al.* 2019; Selmecki, Forche and Berman 2010; Rustchenko-Bulgac 1991) and in genome instability studies (Zhu *et al.* 2012).

Considering the high variability observed in the genome conformation of natural isolates, and its possible relation with genome repeats and instability associated to them,

this project included a laboratory adaptation experiment to determine whether the genome of the natural isolate Y-7124, sequenced in this project, shows signs of instability under temperature stress.

Although temperature stress (37 ºC) produced karyotype modifications in several isolates (12/34), the control conditions at 30 ºC offered more variability (19/36), suggesting that either the genome of Y-7124 is instable even in non-stressful conditions, or that growth in minimal media (SC-Mix) is more stressful than anticipated. Despite the surprising results, Rustchenko (Rustchenko 2007) summarised similar results in *C. albicans,* where the frequency of random chromosome alterations seems to be increased with depletion of nutrients or with reduction of temperature.

Previous studies on *S. cerevisiae* have demonstrated an enrichment on the expression of transposons when grown in minimal media (Halbeisen and Gerber 2009), which might indicate the intention of creating advantageous genetic changes through transposition. Transposition activity has also been related to low temperature stress in *S. cerevisiae* and it is rarely observed at temperatures close to 30 ºC, since the activity of the reverse transcriptase in virus like particles (VLPs) formed during transposition is greatly reduced (Krastanova, Hadzhitodorov and Pesheva 2005). This might explain the higher incidence of modifications at 30 ºC compared to 37 ºC in the adaptation experiment.

Although the presence of several transposons in the genome of the strain Y-7124 has been demonstrated during this project, it is impossible to relate the modifications observed to transposition without studies on the ability of these elements to be expressed and transpose or without the whole genome sequence of the adapted strains.

3.4 CONCLUSIONS

This project has demonstrated through karyotype studies that the genome of natural isolates of *S. stipitis* is plastic and variable. To understand the nature of this variability, two strains commonly used in *S. stipitis* research were compared, NRRL Y-11545, whose genome had been previously sequenced and is used as the reference strain in genetic studies, and NRRL-7124, commonly employed in fermentation research, whose genome was sequenced and assembled in this study by a combination of ONT and Illumina technologies.

Several repetitive regions were detected in both genomes, and these offer variability between the strains, with special emphasis in the transposable elements, content and distrution, subtelomeres and centromeres. The strain Y-11545 offered more repetitive content overall, being the number and distribution of transposable elements the main difference. The transposable elements of *S. stipitis* have been described for the first time in this study. A total of 6 families have been detected, of which 2 non-LTR L1 elements (*Ace* and *Can*), to our knowledge, have been described for the first time in yeasts.

Moreover, the comparison of the whole genome sequence of the two strains allowed the detection of a reciprocal translocation between the chromosome 5 and 7 as the explanation of the differences observed in the karyotype of the strains. This translocation seems associated to the possible instability generated by both the presence of repetitive DNA and transposable elements in the chromosome 7 of the strain Y-11545.

Finally, our attempt to determine through laboratory adaptation how unstable the genome of natural isolates is and how this can condition karyotype modifications also proved that the genome of the natural isolate Y-7124 is prone to modifications, even under non-stressful conditions, reinforcing the plastic nature of the *S. stipitis* genome, previously observed between natural isolates.

3.5 FUTURE WORK

This study has related modifications in the karyotype of natural isolates of *S. stipitis* with a reciprocal translocation potentially associated with repetitive DNA regions in the genome. However, it is based only on the study of the genome sequence of 2 strains. The whole genome sequence of other strains with different chromosome organization might offer a clearer idea on how repetitive regions have conditioned the evolutive history of *S. stipitis.* Moreover, a better understanding on how instability might condition genome modifications and phenotype improvements can be obtained by the genome sequencing and phenotypic characterization of the laboratory adapted strains isolated during this project.

Although several repetitive regions have been detected in this study, and variations are observed in them between the two strains, a deeper understanding is needed.

First, the effects of transposons in evolution in *S. stipitis* is not clear. Only a 15% percent of them seem conserved between the two strains, which indicates that the 85% might still be active and capable of transposition. This can be studied by measuring the mRNA levels of reverse transcriptase by Northern Blot, which will be an indication of transposon expression, and, therefore, their capability to move.

Second, the number of helicases present in the subtelomeres of the two strains is very variable. The repetitive nature of the subtelomeres could cause some errors in the assembly of the reference strain Y-11545, which might have placed the helicases in only a few chromosome ends. The reality of this difference could be easily checked with a PCR strategy, with one primer targeting the helicase DNA sequence, and the other targeting a unique sequence for each subtelomere.

Finally, although the structure of the centromeres has been described before, and variations between strains are detected in this study, the validation of the location is required. This can be achieved by ChIP sequencing techniques using antibodies against centromere specific proteins, such as CENH3, which to our knowledge has never been conducted in *S. stipitis* or any other specie of the centromeric Ty5 cluster.

Chapter 4

Phenotypic diversity in natural isolates of *Scheffersomyces stipitis*

4.1 INTRODUCTION

Yeasts have been historically used for ethanol production. Bioethanol has been proposed as one of the main alternatives to fossil fuels and special attention has been focused on the use of lignocellulosic biomass for fermentations (Mohd Azhar *et al.* 2017). Nevertheless, lignocellulose is a very recalcitrant material and needs to be degraded to release the sugars (commonly a mixture of pentoses and hexoses) that will be used for ethanol fermentation (Robak and Balcerek 2018).

S. cerevisiae, the most used yeast for ethanol fermentation, is not able to ferment pentoses, which means that not all fermentable sugars released are being used and therefore the total amount of ethanol produced is lower than the theoretical maximum. Despite attempts of including pentose fermentation pathways in *S. cerevisiae,* the results obtained are far from being optimal, and secondary by-products such as xylitol, are accumulated (Klimacek *et al.* 2010).

In contrast to *S. cerevisiae, S. stipitis* is able to ferment both types of sugars, and therefore it has great potential for bioethanol production (du Preez and Prior 1985). However, to date, the studies on its suitability for bioethanol production have been based on the use of a few natural isolates (NRRL Y-7124, CBS6054 [or NRRL Y-11545], CBS 5773 [or ATCC 58376], CBS 5774, CBS 5775, CBS 5776 [or ATCC 62970, ATCC 62971 and ATCC 58784 respectively]), on co-cultures with *S. ceverevisiae* or on strains isolated through adaptive laboratory evolution (ALE) experiments (Lee *et al.* 1986; J. P. Delgenes, Moletta and Navarro 1988; Skoog and Hahn-Hägerdal 1990; Amartey and Jeffries 1994; Domínguez *et al.* 2000; Lee *et al.* 2000; Nigam 2001b; De Castro, Oliveira and Furlan 2003; Agbogbo and Wenger 2006; Agbogbo *et al.* 2006; Slininger *et al.* 2006; Agbogbo and Wenger 2007; Agbogbo *et al.* 2008; Slininger, Gorsich and Liu 2009; Silva *et al.* 2011; Slininger *et al.* 2011; Geiger *et al.* 2012; Hao *et al.* 2013; Karagöz and Özkan 2014*;* Slininger *et al.* 2015; Gonçalves, dos Santos and de Macedo 2015; Günan Yücel and Aksu 2015; Augusto Silva De Souza *et al.* 2016; Acevedo, Conejeros and Aroca 2017)

The results presented in this thesis demonstrate that the genome of *S. stipitis* is plastic since natural isolates exhibit different genome organizations. Variations in karyotypes and ploidy levels have been previously reported in *S. stipitis* (Passoth *et al.* 1992) (Talbot and Wayman 1989). However, it is still unknown whether the different genome organization is associated with differences in the phenotype.

Therefore, the aim of this section is to study whether the genome variability observed in natural *S. stipitis* isolates conditions phenotypic traits of interest for ethanol fermentation.

4.2 RESULTS

4.2.1 Phenotypic diversity of *S. stipitis* natural isolates

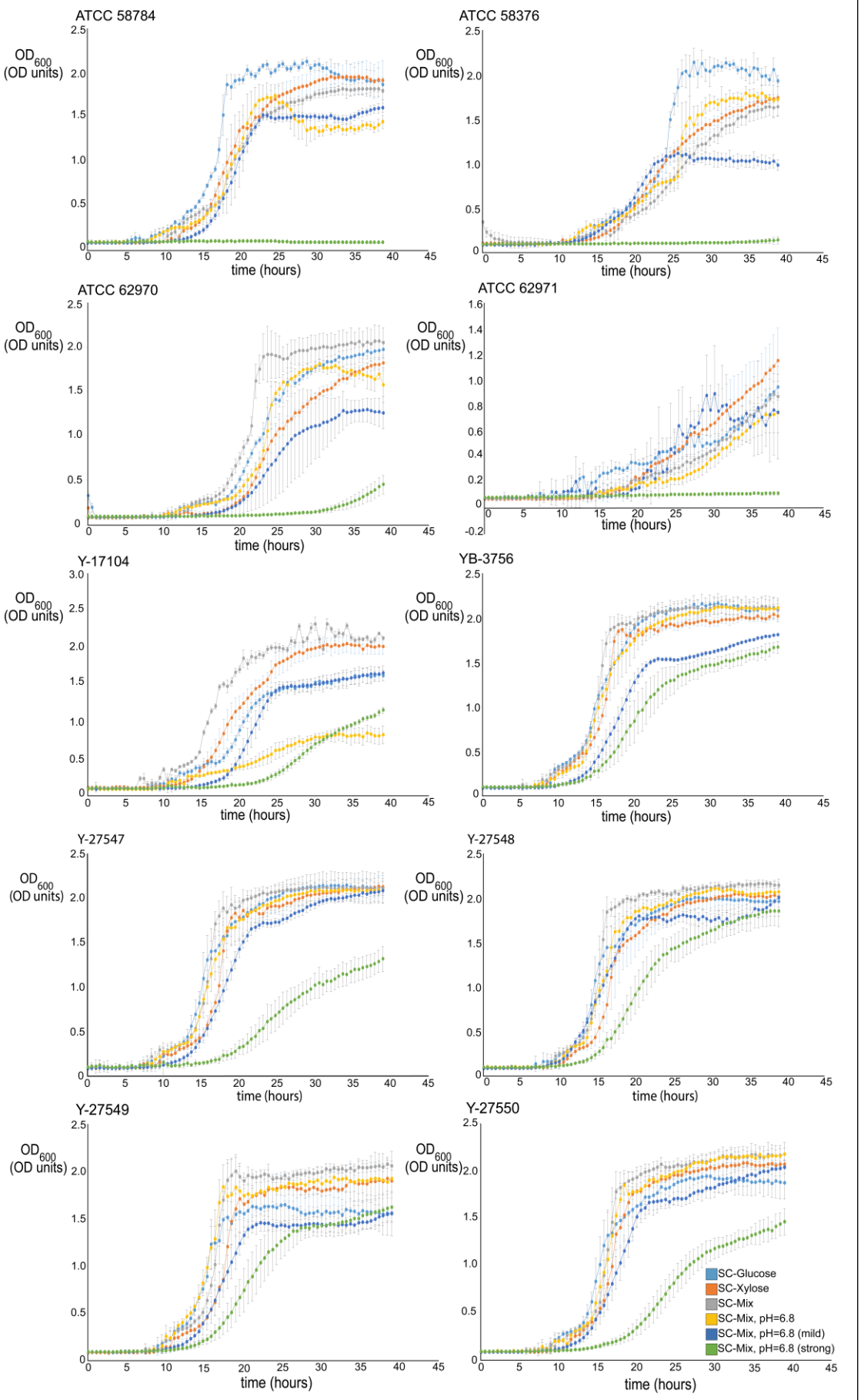
The yeast *S. stipitis* sustains its growth using both five (C5) and six (C6) carbon sugars (such as xylose, ribose [C5], glucose or galactose [C6]) (Pignal 1967). However, despite being the best xylose fermenting yeast described, C6 sugars are preferred for growth and fermentation, which limits the amount of C5 sugars to be consumed (du Preez and Prior 1985; Kilian and van Uden 1988).

Lignocellulose is a glucose and xylose rich bio-source, so the use of strains that are able to consume both of them efficiently is highly recommended (Kim *et al.* 2012). Therefore, the 27 natural isolates were tested for their ability to grow on synthetic media containing different carbon sources: glucose, xylose, and a mixture of both sugars.

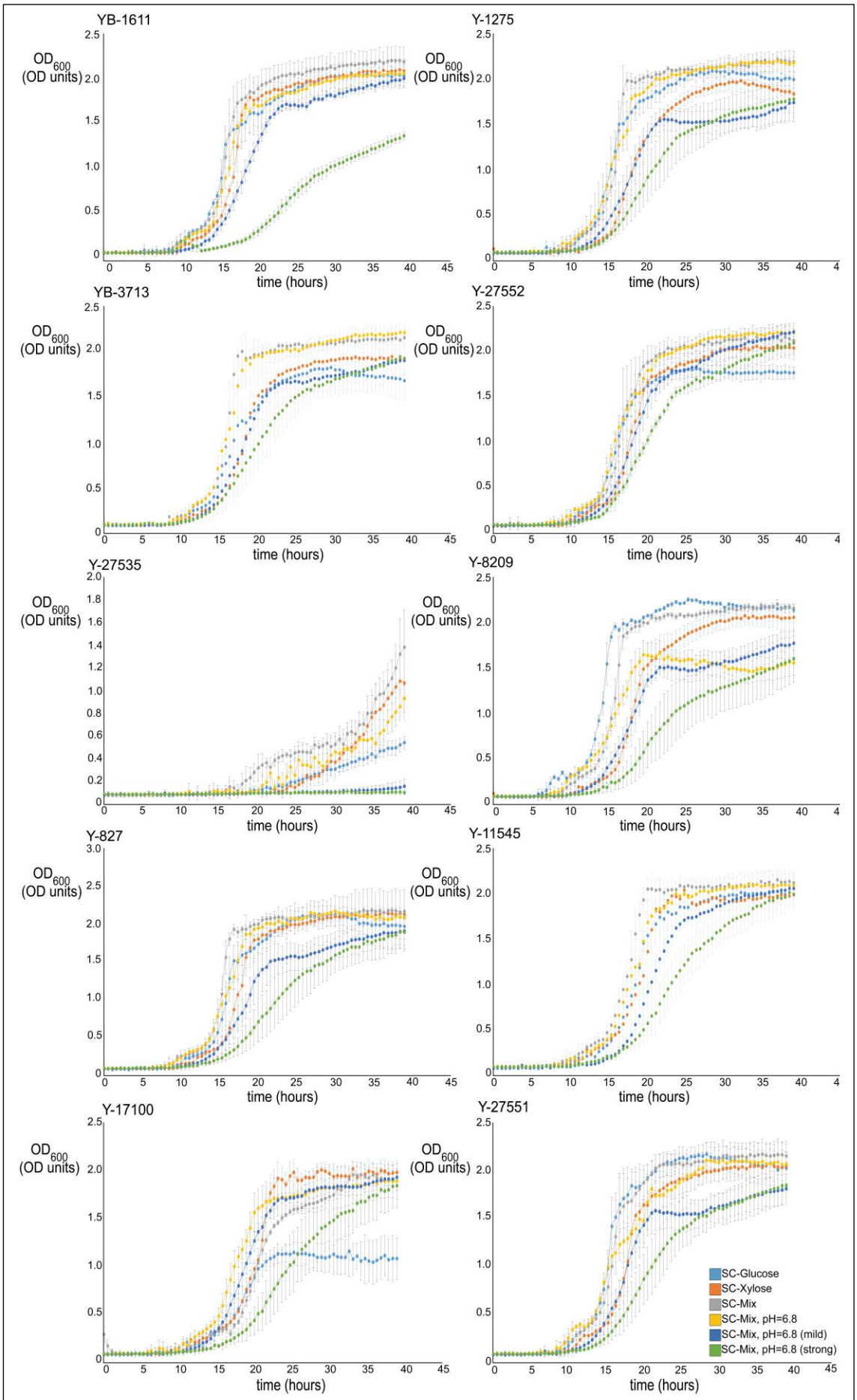
Moreover, the process of hydrolysis of lignocellulosic biomass for the release of fermentable sugars causes also the production of inhibitory compounds that might affect growth and fermentation (Palmqvist and Hahn-Hägerdal 2000; Klinke, Thomsen and Ahring 2004; Ko *et al.* 2015), so finding strains resistant to those stresses is very interesting for industrial purposes. Hence, all the natural isolates were also tested for their ability to grow on synthetic media containing inhibitory compounds commonly found in lignocellulose hydrolysates (Tran and Chambers 1986; Larsson *et al.* 1999; GArcía-Aparicio *et al.* 2006; Slininger *et al.* 2015).

4.2.1.1 Effects of glucose and xylose as carbon sources

The growth of all the natural isolates was measured using SC as media, containing 2% glucose (SC-Glucose), 2% xylose (SC-Xylose) or 1.2% Glucose + 0.8% Xylose (SC-Mix) as carbon sources. The increase in OD_{600} at 30ºC over time was used as a growth indicator (**Figure 4.1**). To understand how the different natural isolates grow in the different medias, three parameters were evaluated as growth indicator: growth rate, maximum OD_{600} and lag time. The growth rate is a measure of the speed at which strains grow, maximum OD_{600} is a measure of productivity of biomass and lag time is a measure of the time that it takes the culture to reach the phase of exponential growth (Mauerhofer *et al.* 2019). All the strains can grow on a media containing glucose, xylose, or a mixture of them (**Figure 4.1**). However, growth rate, maximum OD_{600} and lag time are variable across natural isolates.



(legend at the end of the figure)



(legend at the end of the figure)

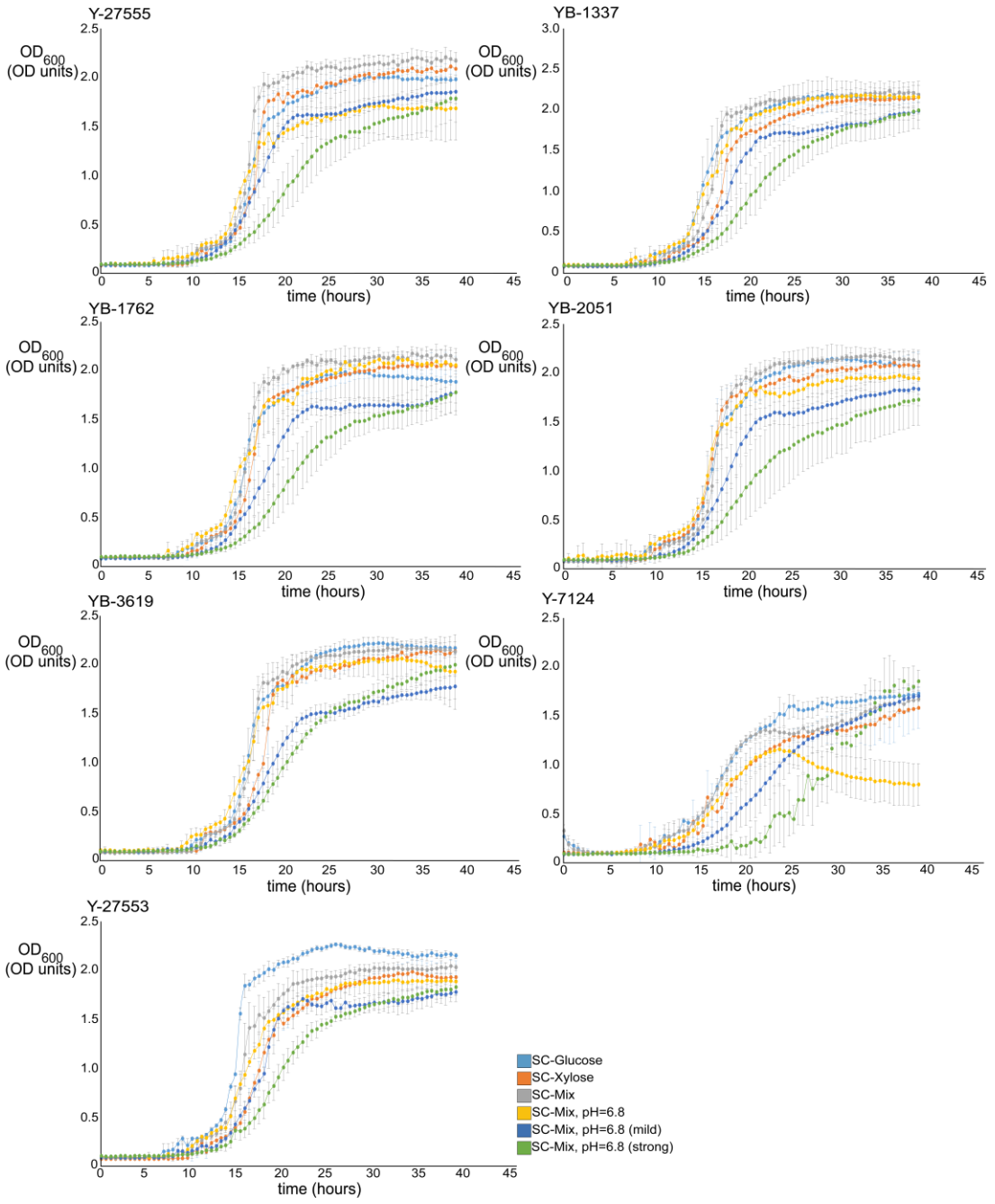


Figure 4.1. Growth curves for *S. stipitis* natural isolates. OD_{600} over time (hours) was measured as growth indicator in media with glucose (light blue), xylose (orange) or a mixture of both (grey) as carbon source. Growth was also studied in the presence of a cocktail of inhibitors commonly found in lignocellulose hydrolysates in two different concentrations: mild (MIN, dark blue) and high (MAX, green) with the pH adjusted to 6.8, and with its respective control (yellow). Experiments were conducted in three technical and three biological replicates.

The complete genome sequence of two of the natural isolates is currently available. The strain Y-11545 was previously sequenced by Jeffries *et al.*(Jeffries *et al.* 2007) and therefore will be used as a the reference strain. Moreover, the strain Y-7124 has been sequenced and assembled to chromosome level during this study (See chapter 3).

The strain Y-11545 has the same growth rate in SC-Glucose and SC-Xylose, however, there is a statistical improvement in SC-Mix (one-way ANOVA, p-value < 0.05). On the other hand, the strain NRRL Y-7124 shows no variations in any of the medias tested. Moreover, when compared, the reference strain Y-11545 and the natural isolate Y-7124 show no statistical differences in the growth rate in glucose or xylose, but the strain Y-7124 grows slower in SC-Mix (one-way ANOVA, p-value < 0.05) (**Figures 4.2-4.4**).

When compared to the reference strain, the isolates Y-27547, Y-27548, Y-12759, Y-27552, Y-8209, Y-8271, Y-27551, YB-1762, YB-3619 and Y-27553 (green) show a statistical improvement on growth rate in SC-Glucose, whereas ATCC 62971, Y-17104, Y-27552 and Y-17100 (red) show a significant decrease. On the other hand, the isolates ATCC 58784, ATCC 58376, ATCC 62970, YB-3756, Y-27549, Y-27550, YB-1611, YB-3713, Y-27555, YB-1337, YB-2051 (grey) show no statistical differences (one-way ANOVA, p-value < 0.05) (**Figure 4.2**).

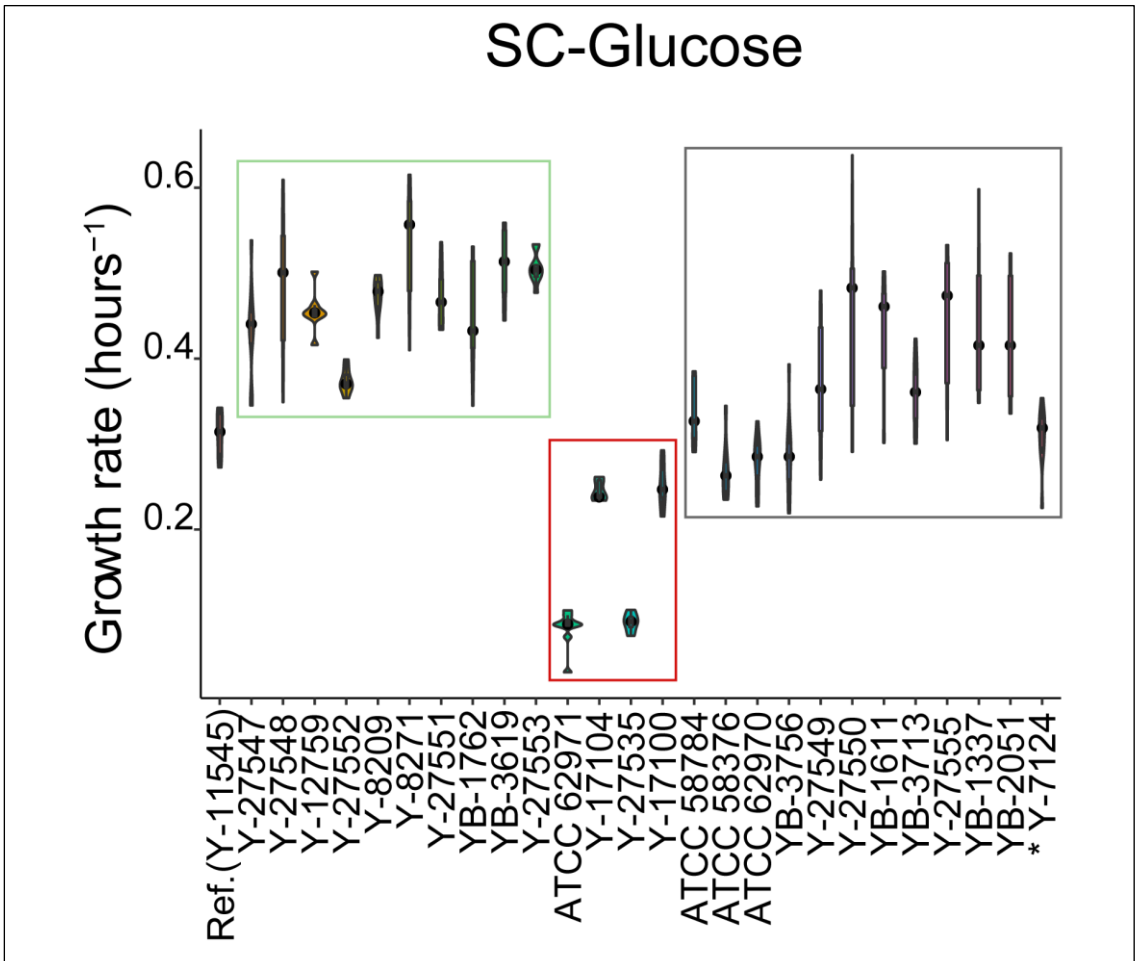


Figure 4.2. Growth rate (hours^{-1}) of *S. stipitis* natural isolates grown in SC-Glucose. Strains inside the green box showed significantly higher growth rates than the reference strain Y-11545, whereas strains in the red box significantly lower. There was no difference respect to the strains present the grey box (one-way ANOVA, p<0.05). Experiments were conducted in three technical and three biological replicates. Ref: Reference strain in genetic studies (Y-11545). * Indicates the strain Y-7124 sequenced during this study.

In SC-Xylose, the strains Y-27547, Y-27548, Y-27549, Y-27550, YB-1611, Y-27552, Y-8209, Y-8271, Y-27555, YB-1337, YB-1762, YB-2051, and YB-3619 (green) show a significantly higher growth rate compared to the reference Y-11545, whereas strains ATCC 62971 and Y-27535 (red) have a significantly lower growth rate. Strains ATCC 58784, ATCC 58376, ATCC 62970, Y-17104, YB-3756, Y-12759, YB-3713, Y-17100, Y-27551 and Y-27553 (grey) have statistically the same growth rate (one-way ANOVA, p-value < 0.05) (**Figure 4.3**).

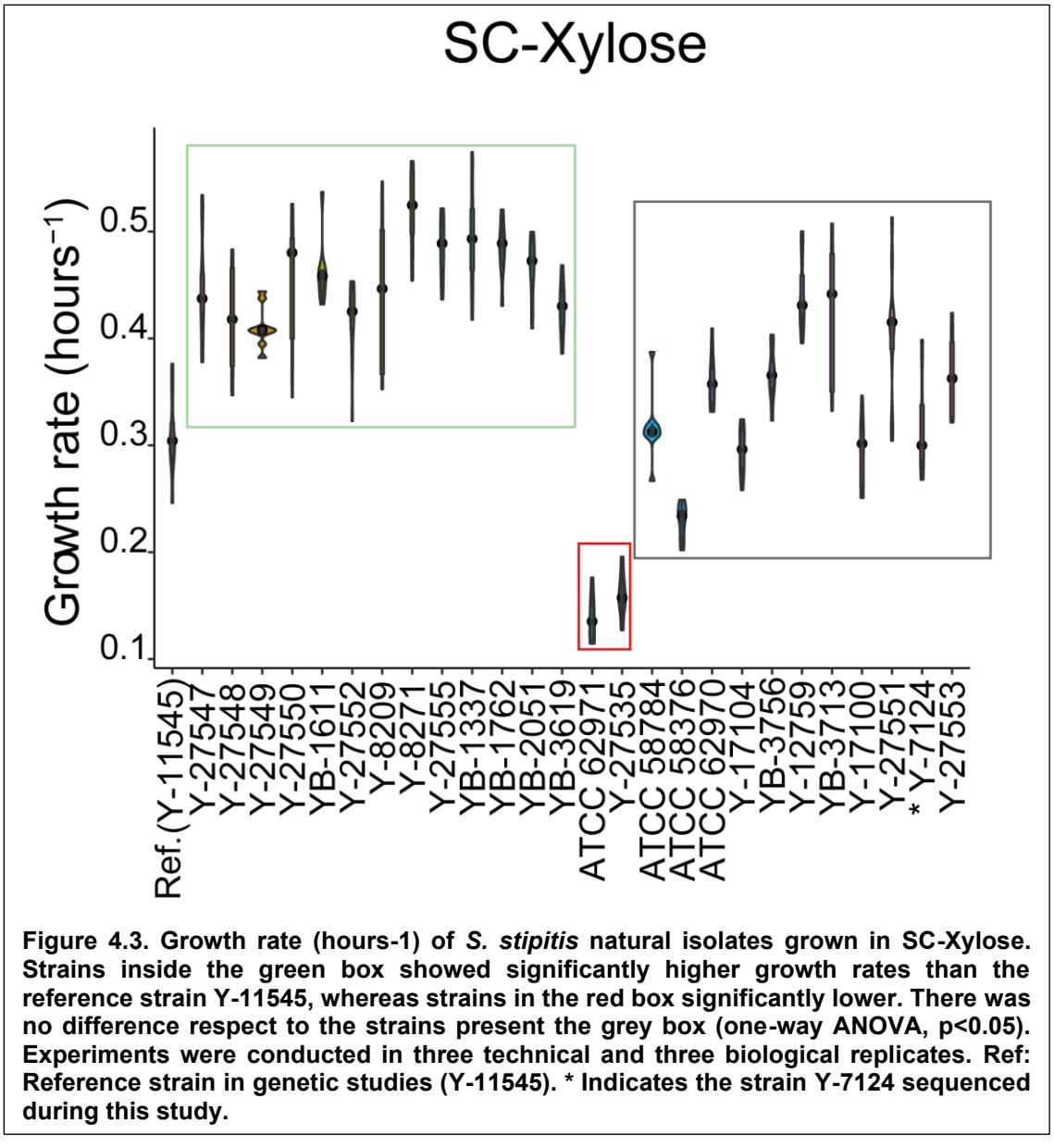


Figure 4.3. Growth rate (hours-1) of *S. stipitis* natural isolates grown in SC-Xylose. Strains inside the green box showed significantly higher growth rates than the reference strain Y-11545, whereas strains in the red box significantly lower. There was no difference respect to the strains present the grey box (one-way ANOVA, p<0.05). Experiments were conducted in three technical and three biological replicates. Ref: Reference strain in genetic studies (Y-11545). * Indicates the strain Y-7124 sequenced during this study.

Lastly, the growth rate in SC-Mix for the strains Y-27547, Y-27548, Y-27549, Y-27550, YB-1611, Y-12759, YB-3713, Y-8209, Y-8271, Y-27555, YB-1337, YB-1762, YB-2051, YB-3619 and Y-27553 (green) is significantly higher compared to the reference strain Y-11545, whereas for strains ATCC 58376, ATCC 62971 and Y-27535 (red box)

it is significantly slower. Finally, the growth rate is statistically the same for ATCC 58784, ATCC 62970, Y-17104, YB-3756, Y-27552, Y-17100 and Y-27551 (grey) (one-way ANOVA, p-value < 0.05) (**Figure 4.4**).

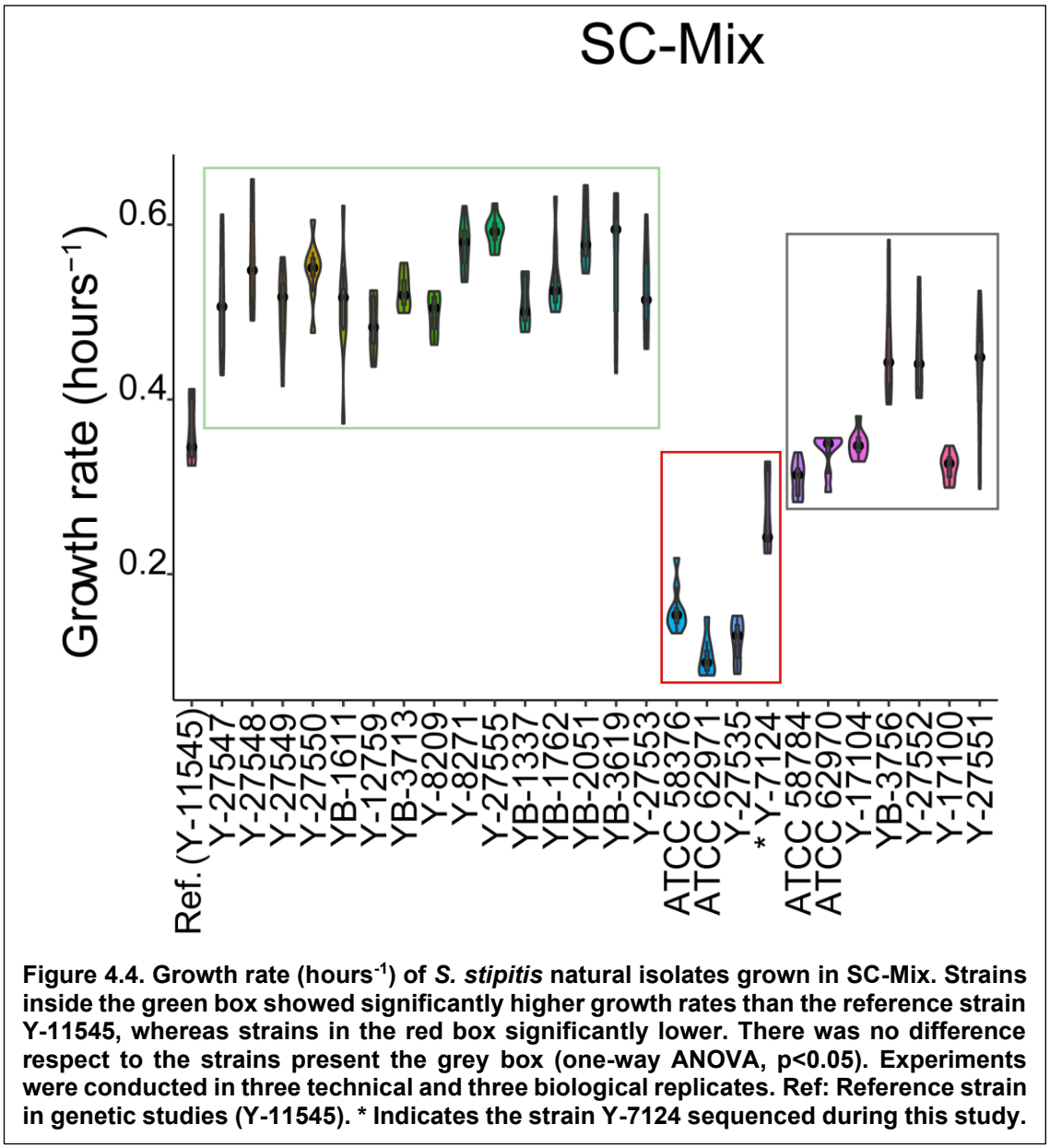


Figure 4.4. Growth rate (hours^{-1}) of *S. stipitis* natural isolates grown in SC-Mix. Strains inside the green box showed significantly higher growth rates than the reference strain Y-11545, whereas strains in the red box significantly lower. There was no difference respect to the strains present the grey box (one-way ANOVA, p<0.05). Experiments were conducted in three technical and three biological replicates. Ref: Reference strain in genetic studies (Y-11545). * Indicates the strain Y-7124 sequenced during this study.

In summary, when the growth rates of natural isolates are compared to the reference strain Y-11545, Y-27547, Y-27548, Y-8209, Y-8271, YB-1762 and YB-3619 show statistical improvements independently of the media, whereas ATCC 62971 and Y-27535 have a consistently slower growth rates.

Maximum OD_{600} can be used as a measurement of biomass productivity in a culture. Both reference strain Y-11545 and the strain Y-7124 show no preference for any of the sugar carbon combinations, since there are no significant differences between SC-Glucose, SC-Xylose and SC-Mix (one-way ANOVA, p-value < 0.05). However, when compared between them, natural isolate Y-7124 showed a significant decrease

compared to Y-11545 in SC-Glucose and SC-Mix (from OD_{600} 2.06±0.11 to 1.77±0.06 and from 2.23±0.11 1.70±0.09).

None of the other natural isolates reached higher OD_{600} than the strain Y-11545 regardless of the growth media studied (**Figure 4.5, A-C**). In contrast, the strains ATCC 62971 and Y-27535 show both a significant decrease in the maximum growth (**Figure 4.5, A-C**) in all the media (red) (one-way ANOVA, p-value < 0.05).

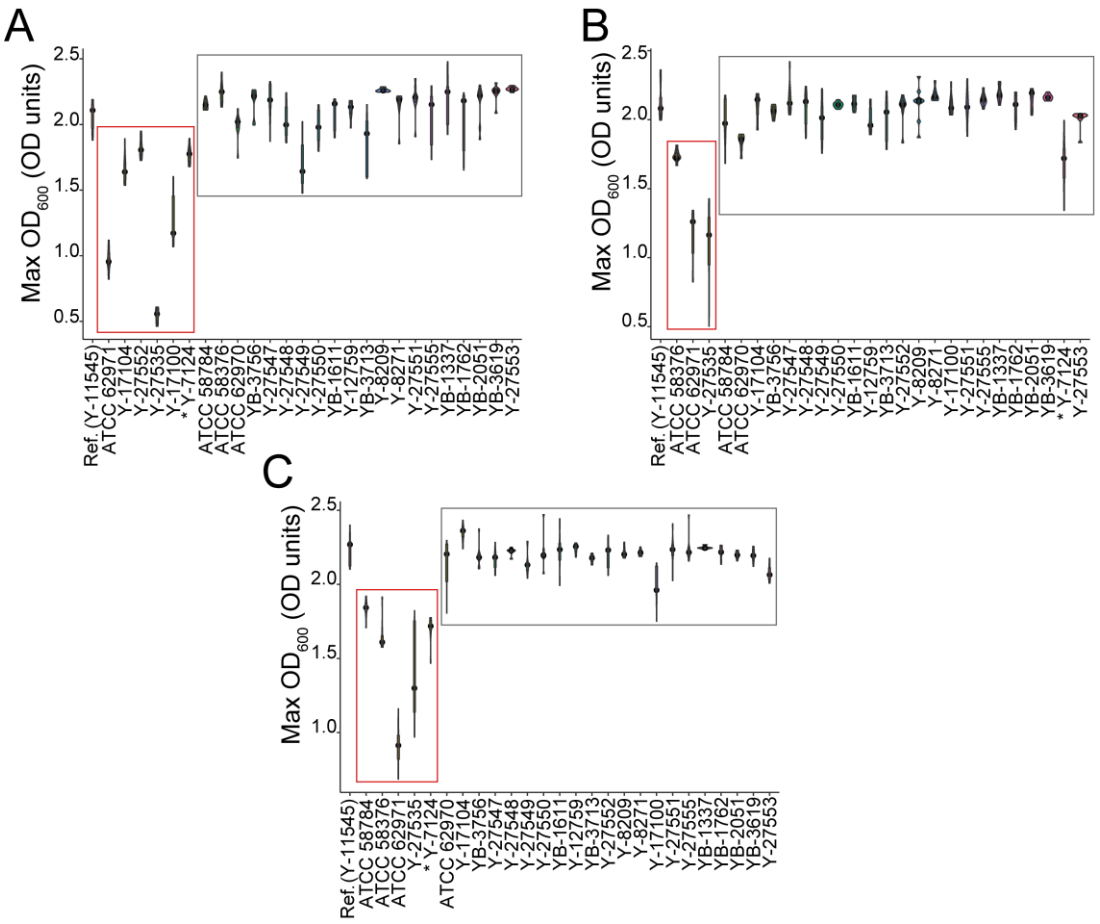


Figure 4.5. Maximum OD_{600} (OD units) of *S. stipitis* natural isolates grown in SC-Glucose (A), SC-Xylose (B) and SC-Mixture (C). Strains inside the green box showed significantly higher maximum OD_{600} than the reference strain Y-11545, whereas strains in the red box significantly lower. There was no difference respect to the strains present the grey box (one-way ANOVA, p<0.05). Experiments were conducted in three technical and three biological replicates. Ref: Reference strain in genetic studies (Y-11545). * Indicates the strain Y-7124 sequenced during this study.

Natural isolate ATCC 58376 grew significantly less than the reference strain in SC-Xylose and SC-Mix. Natural isolate ATCC 58784 grew significantly less than Y-11545 in SC-Mix, and finally, natural isolates Y-17104, Y-27552 and Y-17100 grew significantly less than Y-11545 in SC-Glucose (one-way ANOVA, p-value < 0.05) (**Figure 4.5**).

In summary, most of the strain show similar productivity to the reference strain Y-11545 in all three sugar carbon sources combination, whereas only the natural isolates ATCC 62971 and Y-27535 show lower maximum OD_{600} in the three media.

The lag time, a measure of the time that the culture requires to reach the phase of exponential growth, is the most variable parameter in Y-11545. The shortest lag time (796.2±21.1 minutes) is found on SC-Glucose, whereas the longest lag time (958.6±33.1 minutes) is found in SC-Xylose. The shortest lag time is also exhibited in SC-Glucose in the strain Y-7124 (673.3±54.8 minutes), but in this case there are no statistical variations between SC-Xylose and SC-Mix. When compared, the strain Y-7124 exhibits shorter lag time in SC-Glucose (being the only strain with shorter lag time than Y-11545), whereas it is maintained in the other media (one-way ANOVA, p-value < 0.05)

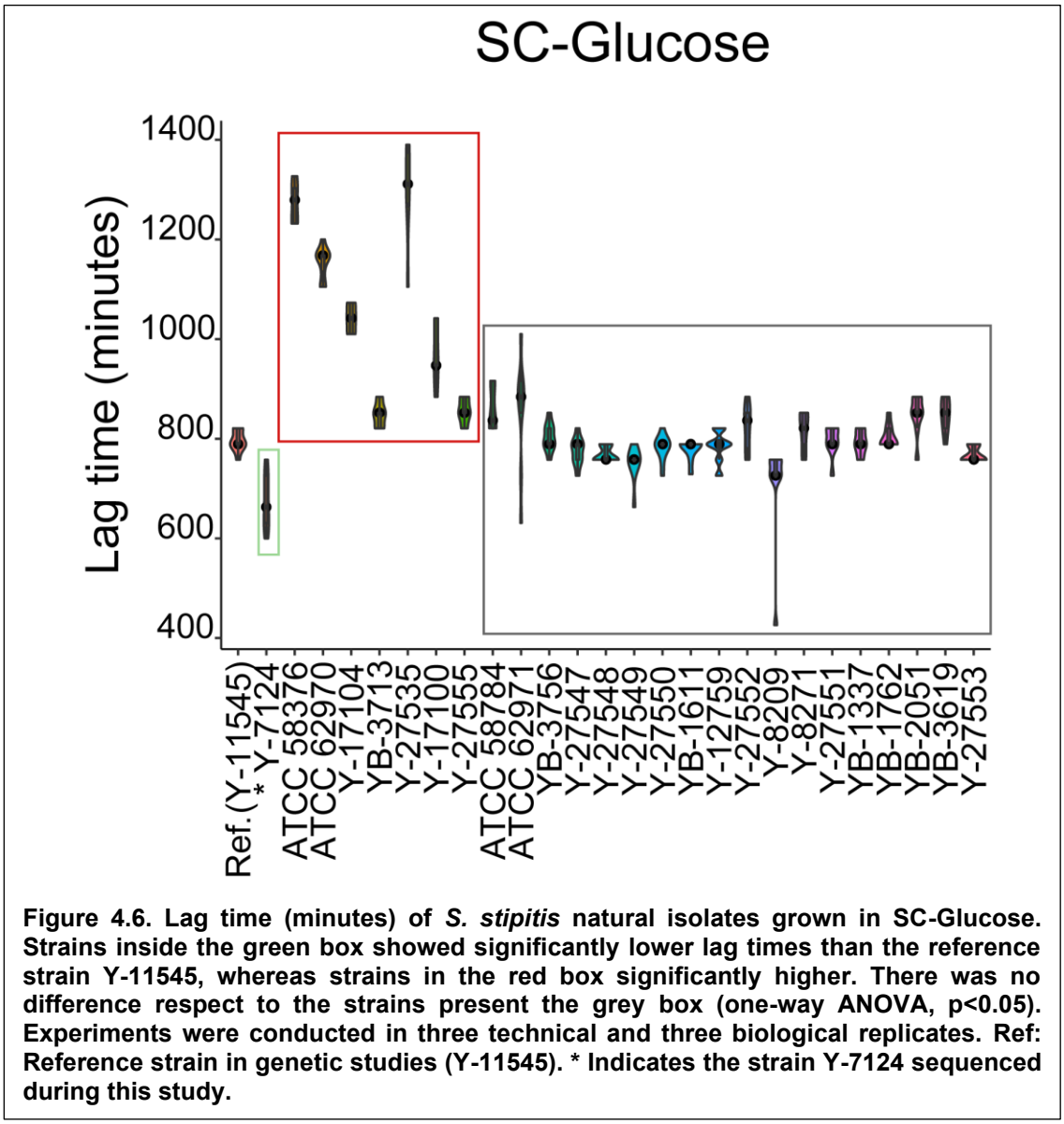


Figure 4.6. Lag time (minutes) of *S. stipitis* natural isolates grown in SC-Glucose. Strains inside the green box showed significantly lower lag times than the reference strain Y-11545, whereas strains in the red box significantly higher. There was no difference respect to the strains present the grey box (one-way ANOVA, p<0.05). Experiments were conducted in three technical and three biological replicates. Ref: Reference strain in genetic studies (Y-11545). * Indicates the strain Y-7124 sequenced during this study.

Several strains show longer lag time in SC-Glucose when compared to the strain Y-11545 (ATCC 62970, Y-27535 and Y-17100) (one-way ANOVA, p-value 0.05) (**Figure 4.6**).

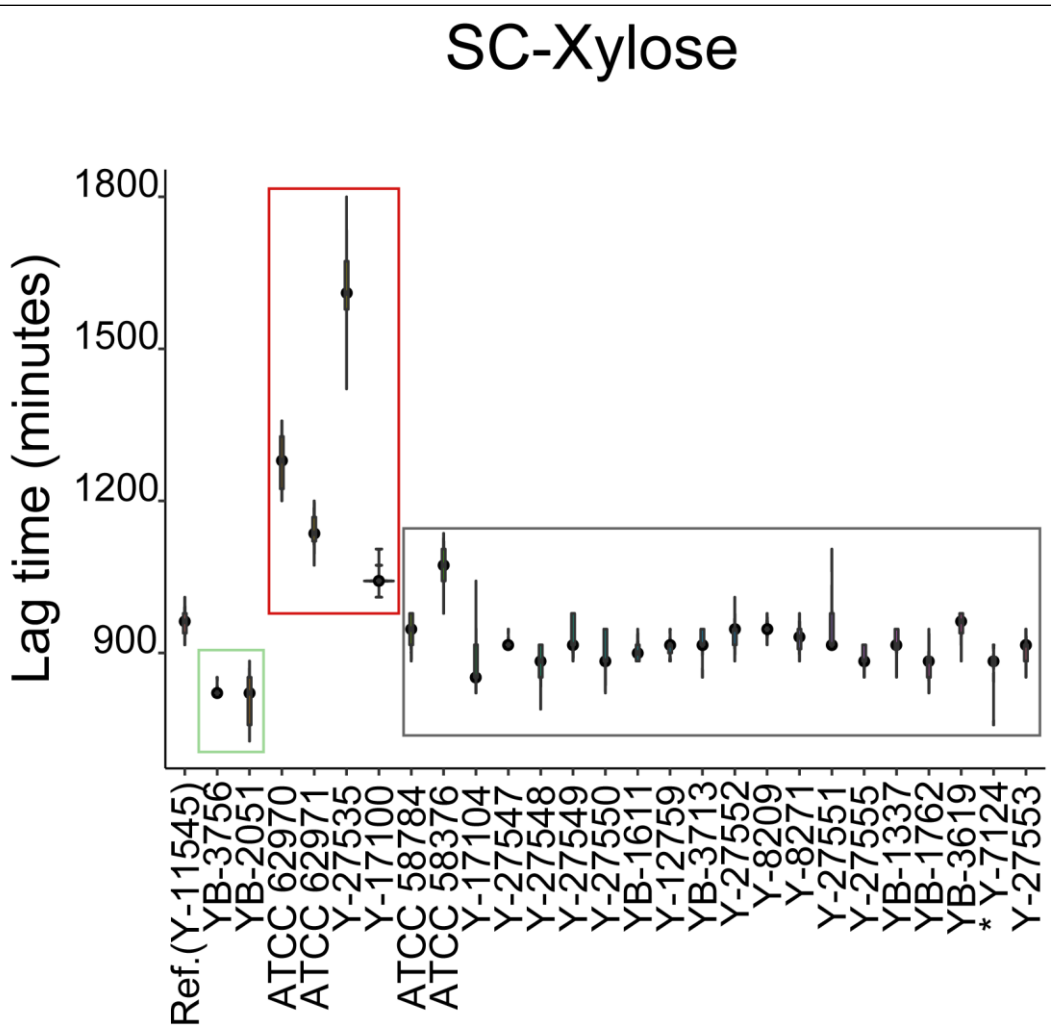


Figure 4.7. Lag time (minutes) of *S. stipitis* natural isolates grown in SC-Xylose. Strains inside the green box showed significantly lower lag times than the reference strain Y-11545, whereas strains in the red box significantly higher. There was no difference respect to the strains present the grey box (one-way ANOVA, p<0.05). Experiments were conducted in three technical and three biological replicates. Ref: Reference strain in genetic studies (Y-11545). * Indicates the strain Y-7124 sequenced during this study.

In SC-Xylose, natural isolates YB-3756 and YB-2051 have a significantly shorter lag time compared to the reference strain, while it was significantly longer for natural isolates ATCC 62970, Y-27535, Y-17100 and ATCC 62971 (one-way ANOVA, p-value < 0.05) (**Figure 4.7**).

Finally, none of the natural isolates showed shorter lag time than the reference strain Y-11545 for SC-Mix, but in the case of ATCC 58376, ATCC 62970, ATCC 62971, Y-27535 and Y-17100 it was significantly longer (**Figure 4.8**).

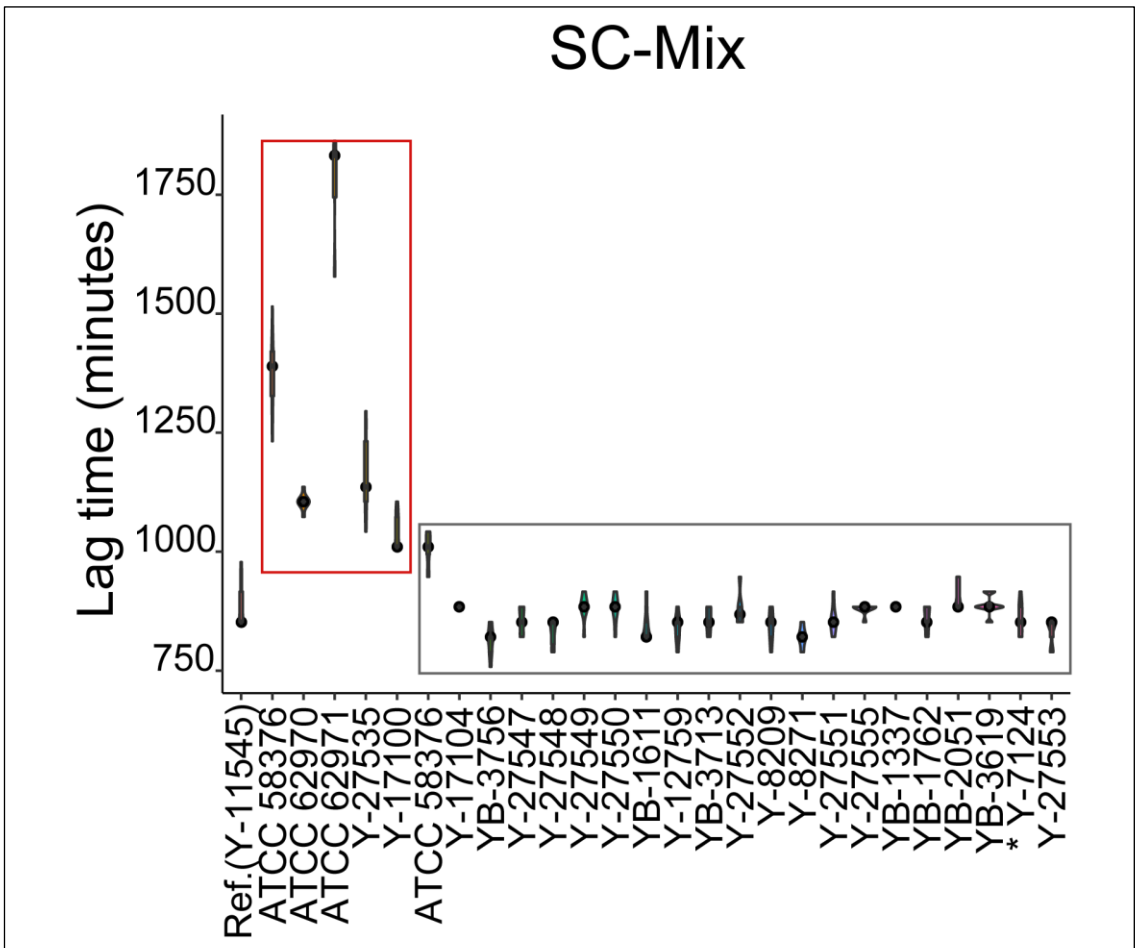


Figure 4.8. Lag time (minutes) of *S. stipitis* natural isolates grown in SC-Mix. Strains inside the green box showed significantly lower lag times than the reference strain Y-11545, whereas strains in the red box significantly higher. There was no difference respect to the strains present the grey box (one-way ANOVA, $p<0.05$). Experiments were conducted in three technical and three biological replicates. Ref: Reference strain in genetic studies (Y-11545). * Indicates the strain Y-7124 sequenced during this study.

In summary, the majority of the strains show statistically the same lag time than the reference strain Y-11545, whereas the natural isolates ATCC 58376, ATCC 62970 and Y-27535 show longer lag phase in all three media.

Bioethanol production at industry level from yeasts require strains that among other characteristics, have a rapid adaptation to the growing conditions (which is directly indicated by short lag time (Swinnen 2004)), and high growth rate and biomass productivity (Mohd Azhar *et al.* 2017).

According to this, five natural isolates can be selected as best performers in SC-Glucose: Y-27547, Y-27548, YB-1611, Y-8209 and Y-27553, since they exhibit simultaneously the statistical maximum growth rate, maximum OD_{600}, and minimum lag time. Similarly, ten strains show simultaneously the best parameters for SC-Xylose: Y-27547, Y-27548, Y-27550, YB-1611, Y-12759, YB-3713, Y-27555, YB-1337, YB-1762 and YB-2051. Lastly, five natural isolates exhibited the highest growth rate and maximum

OD_{600} and minimum lag time simultaneously in SC-Mix: Y-27548, YB-1611, YB-3713, Y-8271 and Y-27553.

Hence, the strains that exhibit the best growth parameters regardless of the media are Y-27548 and YB-1611.

Conversely, the natural isolates ATCC 62971 and Y-27535 seem to be the strains that always exhibit the worst growth parameters for all the three media studied.

4.2.1.2 Effects of growth inhibitors

The production of bioethanol from microbes demands the accessibility of fermentable sugars that can be metabolised. Carbohydrates required for both yeast growth and fermentation from lignocellulosic biomass are released by different hydrolysis strategies (Lange 2007).

However, during this process other compounds are also produced, and these can often be inhibitory of both biomass and ethanol production (Palmqvist and Hahn-Hägerdal 2000; Klinke, Thomsen and Ahring 2004; Ko *et al.* 2015). An ideal bioethanol producer strain should be able to grow in the presence of these inhibitory compounds.

Therefore, *S. stipitis* natural isolates were tested for their ability to grow in liquid media containing an inhibitory compounds cocktail at two different concentrations (mild and strong). The inhibition cocktail was prepared with some of the inhibitory compounds commonly found in lignocellulose hydrolysates at its minimum (mild) and maximum concentrations (strong) (See Chapter 2) (Tran and Chambers 1986; Larsson *et al.* 1999; GArcía-Aparicio *et al.* 2006; Slininger *et al.* 2015). The experiment was conducted in SC-Mix, since a mixture of glucose and xylose is expected in lignocellulose hydrolysates, and OD_{600} at 30 ºC was measured over time as a growth indicator. Natural isolates were also grown in SC-Mix liquid media with a pH of 6.8±0.1 as a control. This was necessary to compensate the effects of acetic acid, one of the inhibitory compounds of the media.

The presence of inhibitory compounds has distinct effects in the different natural isolates of *S. stipitis*. The natural isolate Y-27535 showed the highest sensitivity to inhibitory conditions, since it was not able to grow in any of the concentrations studied. Moreover, ATCC 58784, ATCC 58376 and ATCC 62971 showed high sensitivity at strong inhibitory conditions, although they were able to grow when the conditions were mild.

None of the natural isolates that were able to grow in the presence of inhibitors showed better resistance behaviour than the reference strain Y-11545, since the comparison of the ratios for both mild and strong inhibitory conditions did not show a significant difference for any of the parameters studied (**Figures 4.9, 4.10 and 4.11**).

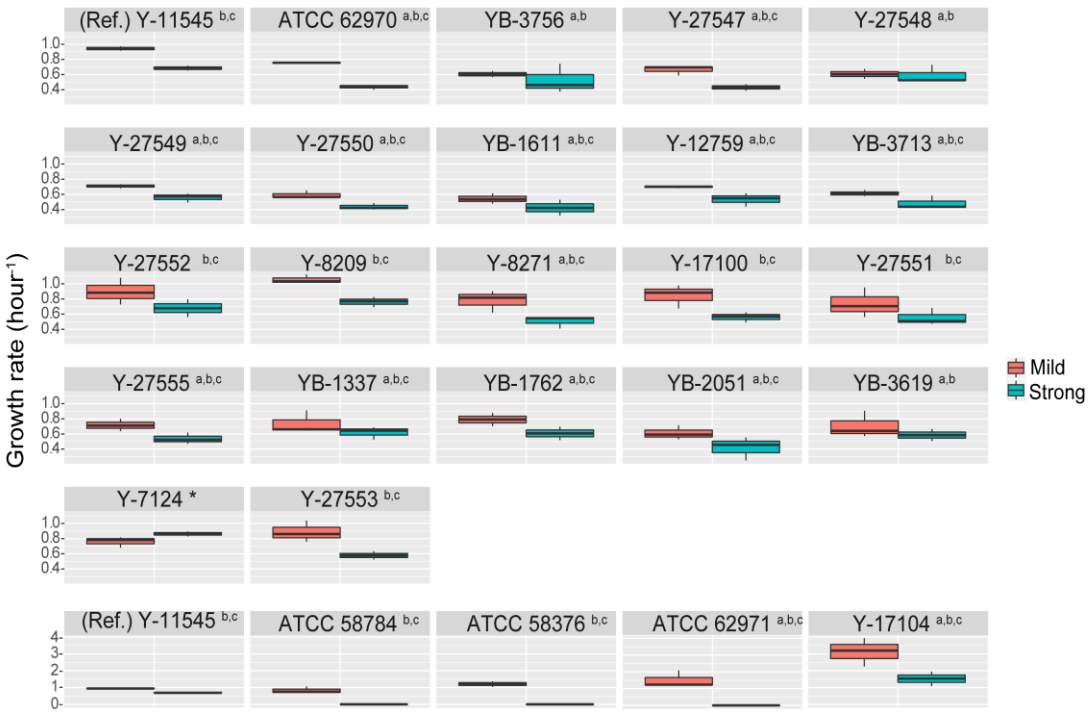


Figure 4.9. Ratios of the growth rate (hours^{-1}) in inhibitory conditions and control media for each natural isolate of *S. stipitis*. Abbreviations: a: Statistical difference between control and mild inhibitory conditions. b: Statistical difference between control and strong inhibitory conditions c: Statistical difference between mild and strong inhibitory conditions. Ref: Reference strain in genetic studies (Y-11545). * Indicates the strain Y-7124 sequenced during this study.

Mild inhibitory conditions do not cause an important decrease on the growth rate of the reference strain Y-11545 respect to the control, but it is significantly decreased in strong inhibitory conditions (ratio of 0.684±0.032). On the other side, the natural isolate Y-7124 was the only strain that did not suffer a decrease in the growth rate in the presence of any of the inhibitor concentrations tested (**Figure 4.9**).

Mild inhibitory conditions did not affect significantly the growth rate of the strains ATCC 58784, ATCC 58376, Y-27552, Y-8209, Y-11545, Y-17100, Y-27551 and Y-27553 (with relative growth rate ratios close to 1), however, these strains suffered a significant decrease in their growth rate when grown under strong inhibitory conditions (**Figure 4.9**).

Conversely, the strains ATCC 62970, Y-27547, Y-27549, Y-27550, YB-1611, Y-12759, YB-3713, Y-8271, Y-27555, YB-1337, YB-1762 and YB-2051 showed a significant decrease in the growth rate when grown in mild inhibitory conditions, and this decrease was more pronounced under strong inhibition (**Figure 4.9**).

Lastly, the strains YB-3756, Y-27548 and YB-3619 showed sensitivity to inhibitors, with a significant decrease in the growth rate, but there was no statistical difference between mild and strong inhibitory conditions (**Figure 4.9**).

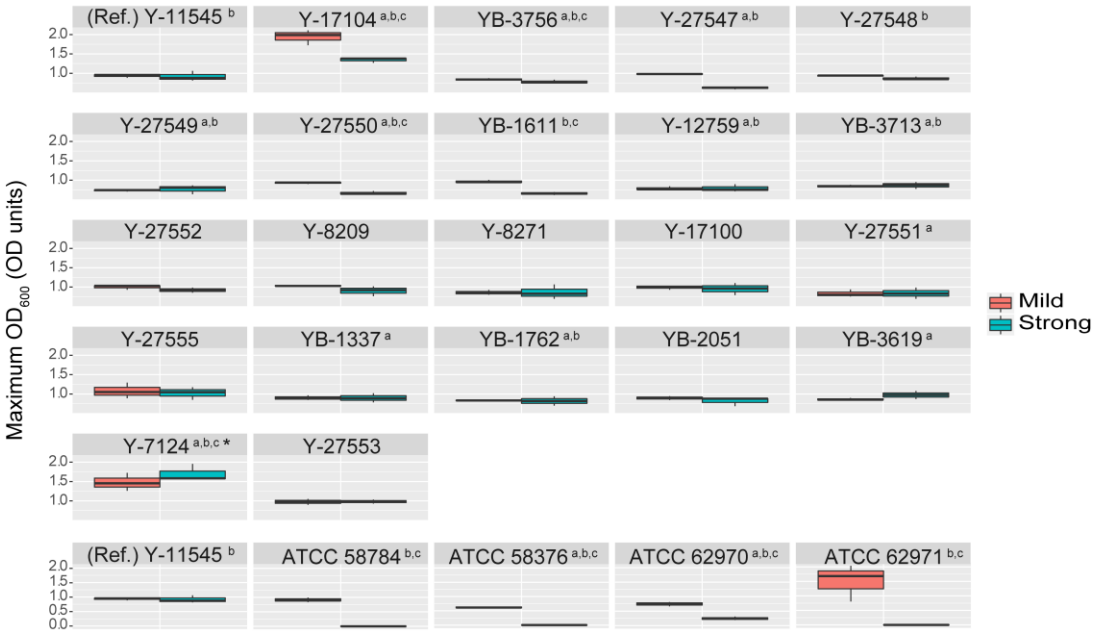


Figure 4.10. Ratios of the maximum OD_{600} (OD units) in inhibitory conditions and control media for each natural isolate of *S. stipitis*. Abbreviations: a: Statistical difference between control and mild inhibitory conditions. b: Statistical difference between control and strong inhibitory conditions c: Statistical difference between mild and strong inhibitory conditions. Ref: Reference strain in genetic studies (Y-11545). * Indicates the strain Y-7124 sequenced during this study.

Mild inhibitory conditions did not significantly affect the maximum OD_{600} of the reference strain Y-11545, although it was lower for strong inhibitory conditions. However, the difference between mild and strong inhibition was not significant. The same effect was observed for the strain Y-27548. Similarly, the natural isolates Y-27551, YB-1337 and YB-3619 did not show any significant difference between the challenging conditions, but surprisingly, mild inhibition was statistically lower than the control, but strong inhibition was not (**Figure 4.10**).

Surprisingly, the strain Y-7124 increased significantly its maximum OD_{600} in inhibitory conditions, reaching the maximum under strong inhibition. However, the productivity was still lower than in the reference strain Y-11545 in mild inhibitory conditions, although it was the same in strong inhibition (one-way ANOVA $p < 0.05$).

None of the inhibitory conditions tested affected significantly the maximum OD_{600} for the strains Y-27552, Y-8209, Y-8271, Y-17100, Y-27555, YB-2051, Y-27553. Moreover, mild inhibitory conditions did not affect it for natural isolates ATCC 58784, Y-27547, YB-1611, but it was statistically lower for strong inhibitory conditions (**Figure 4.10**).

Natural isolate Y-27550 was affected significantly by mild inhibitory conditions and this was even more pronounced for strong inhibition. Conversely, although the

natural isolates Y-27549, Y-12759, YB-3713, Y-27535 and YB-1762 showed that the maximum OD_{600} was lower when grown under inhibitory conditions, there was no statistical difference between both challenging conditions (**Figure 4.10**).

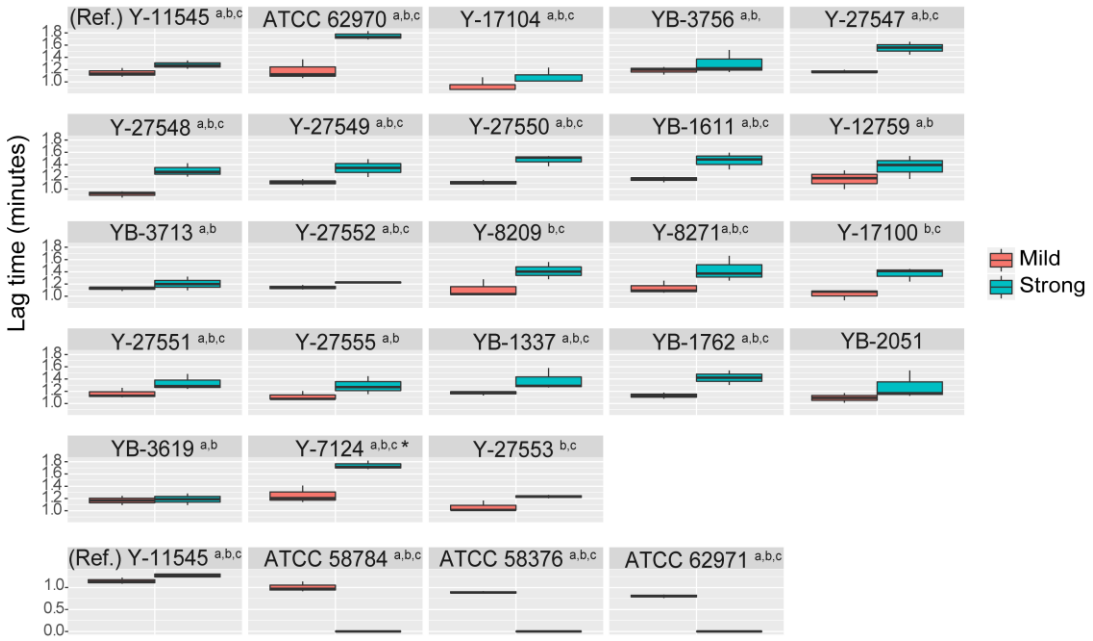


Figure 4.11. Ratios of the lag time (minutes) in inhibitory conditions and control media for each natural isolate of *S. stipitis*. Abbreviations: a: Statistical difference between control and mild inhibitory conditions. b: Statistical difference between control and strong inhibitory conditions c: Statistical difference between mild and strong inhibitory conditions. Ref: Reference strain in genetic studies (Y-11545). * Indicates the strain Y-7124 sequenced during this study.

Lag time was the most affected parameter by inhibitory conditions. Both the reference strain Y-11545 and the strain Y-7124 suffered a significant delay in growth under mild inhibitory conditions, and this delay was more pronounced under strong inhibitory conditions (**Figure 4.11**), with the strain Y-7124 suffering a more pronounced delay under strong inhibition when compared to Y-11545 (one way ANOVA $p < 0.05$).

Natural isolate YB-2051 was the only strain whose lag time was not significantly affected by inhibitors. Mild inhibitory conditions did not affect the lag time of ATCC 58784, Y-27548, Y-8209, Y-17100 and Y-27553, but it increased when they were grown in strong inhibitory conditions (**Figure 4.11**).

Natural isolates ATCC 62970, Y-27547, Y-27549, Y-27550, YB-1611, Y-27552, Y-8271, Y-11545, Y-27551, YB-1337, and YB-1762 show an increasing delay in the lag time as the concentration of inhibitors increase. Similarly, in strains YB-3756, Y-12759, YB-3713, Y-27555 and YB-3619 there is also a significant increase in the lag time when inhibitors are present, but there is no statistical difference between strong and mild conditions (**Figure 4.11**).

Three natural isolates offer a surprising result for all the parameters studied. Both ATCC 62971 and Y-17104 show higher growth rate when grown with inhibitors, although ATCC 62971 is completely inhibited under strong conditions (**Figure 4.9**). Y-17104 was the only natural isolate that showed shorter lag time in the presence of inhibitors (mild conditions), although it was significantly longer when grown in high concentration of inhibitors (**Figure 4.11**). Moreover, it also showed higher maximum OD_{600} in the presence of inhibitors, with the maximum reached in mild conditions (**Figure 4.10**), similarly to the effect observed in Y-7124.

In conclusion, the natural isolate Y-27535 showed the highest sensitivity to inhibitors, since it was not able to grow in none of the conditions studied. Contrarily, Y-27553 and YB-2051 showed the best resistance phenotype, since their growth rate and lag time were not affected under mild inhibitory conditions, and their maximum OD_{600} were not affected, although Y-27553 has a higher growth rate in both inhibitory conditions. The improvements observed in Y-17104 and Y-7124 are also remarkable, although their growth rates are inferior to Y-27553.

4.2.3 Biofilm related phenotypes

Biofilm reactors have been proven to enhance the production of certain value added products, such as ethanol, organic acids, antibiotics, enzymes or polysaccharides, since they can increase volumetric productivity by maintaining high biomass concentration in the reactors (A Demirci, K-LG Ho and AL Pometto III 1997). A higher tolerance to ethanol has also been observed in yeast in fixed bed biofilm reactors (Cheng, Demirci and Catchmark 2010), which might indicate that strains with biofilm forming properties could be more suitable for industrial fermentations. Then, the aim of this section was to determine if natural isolates of *S. stipitis* show biofilm forming characteristics that could be potentially interesting in bioethanol industry.

To determine the diversity of phenotypes present in natural isolates of *S. stipitis*, two different biofilm related phenotypes were studied: agar invasion and cell sedimentation (Hope and Dunham 2014). The agar invasion assay measures the ability of a strain to invade solid agar media, and it is quantified by the density of cells remaining in the plate when the colonies are washed off the surface (See Chapter 2). The cell sedimentation assays give an idea on how a disperse population aggregates in a liquid culture system, and it is important for bioengineering process design.

As previously stated in section 4.2.1, different growth is observed for each strain and media studied. To avoid these variations having an effect in the phenotypes studied the sedimentation profiles were conducted starting with the same cell concentration (OD_{600}=1.0), and the agar invasion was normalised calculating the ratio of the cell

invasion of the agar versus the cell growth on the surface of the plate. All experiments were conducted in SC-Mix media.

Natural isolates differ in their ability to invade SC-Mix Agar plates (**Figure 4.12**).

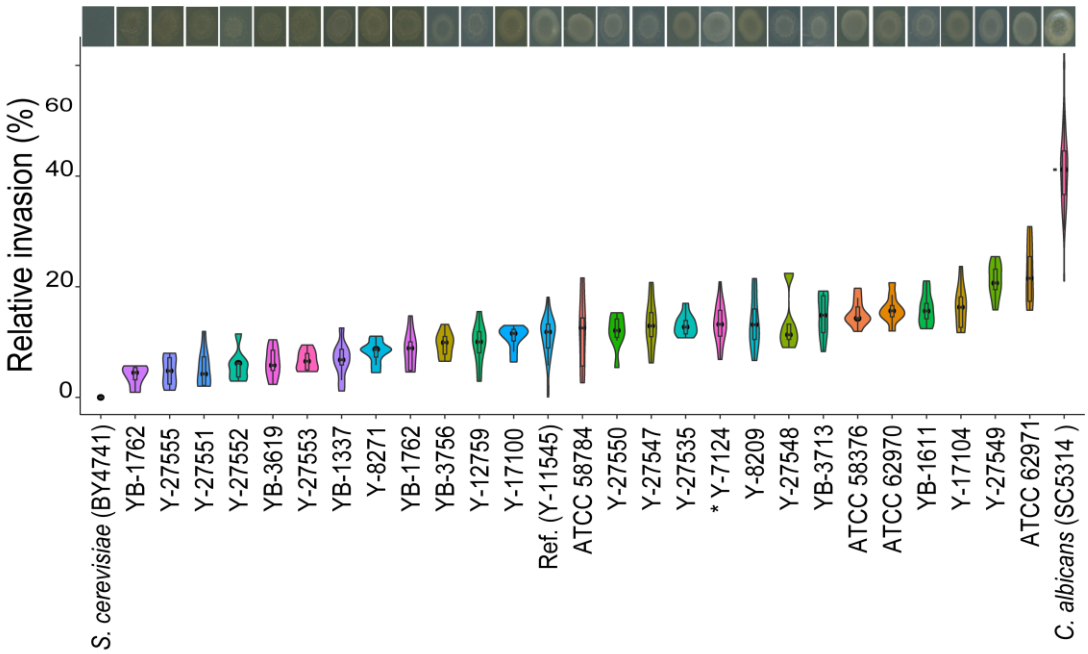


Figure 4.12. Relative agar invasion (%) of natural isolates of *S. stipitis* grown in SC-Mix. Experiments were conducted growing the strains in SC-Mix agar plates for 5 days at 30 ºC. The plates were then photographed before and after washing them with water and the grey intensity of the growth over and inside the agar was compared. *Candida albicans SC5314* was added as a reference for positive agar invasion, whereas *S. cerevisiae BY4741* was added as a control for no agar invasion. Experiments were conducted in three technical and three biological replicates. Ref: Reference strain in genetic studies (Y-11545). * Indicates the strain Y-7124 sequenced during this study.

The strains *S.cerevisiae* BY4741 and *C. albicans* SC5314 were used respectively as negative and positive control of agar invasion.

All *S. stipitis* strains showed agar invasion when compared to *S. cerevisiae*, although none of them was as invasive as *C. albican*s (41.31±7.73%), a yeast able to switch from yeast to hyphal growth (Slutsky, Buffo and Soll 1985; Pomés, Gil and Nombela 1985; Slutsky *et al.* 1987). The strain ATCC 62971 showed the highest relative adhesion (22.55±5.63%), whereas YB-1762 the lowest (3.94±1.94%). Natural isolates ATCC 62971, Y-27549 and Y-17104 showed significant improvements in adhesion respect to the reference strain Y-11545 (2.03, 1.91 and 1.48 folds difference), whereas natural isolates YB-1762, Y-27555 and Y-27551 showed statistical less invasion than the reference strain Y-11545 (0.36, 0.44 and 0.46 fold difference). The strains Y-7124 and Y-11545 were equally invasive (one-way ANOVA, $p<0.05$).

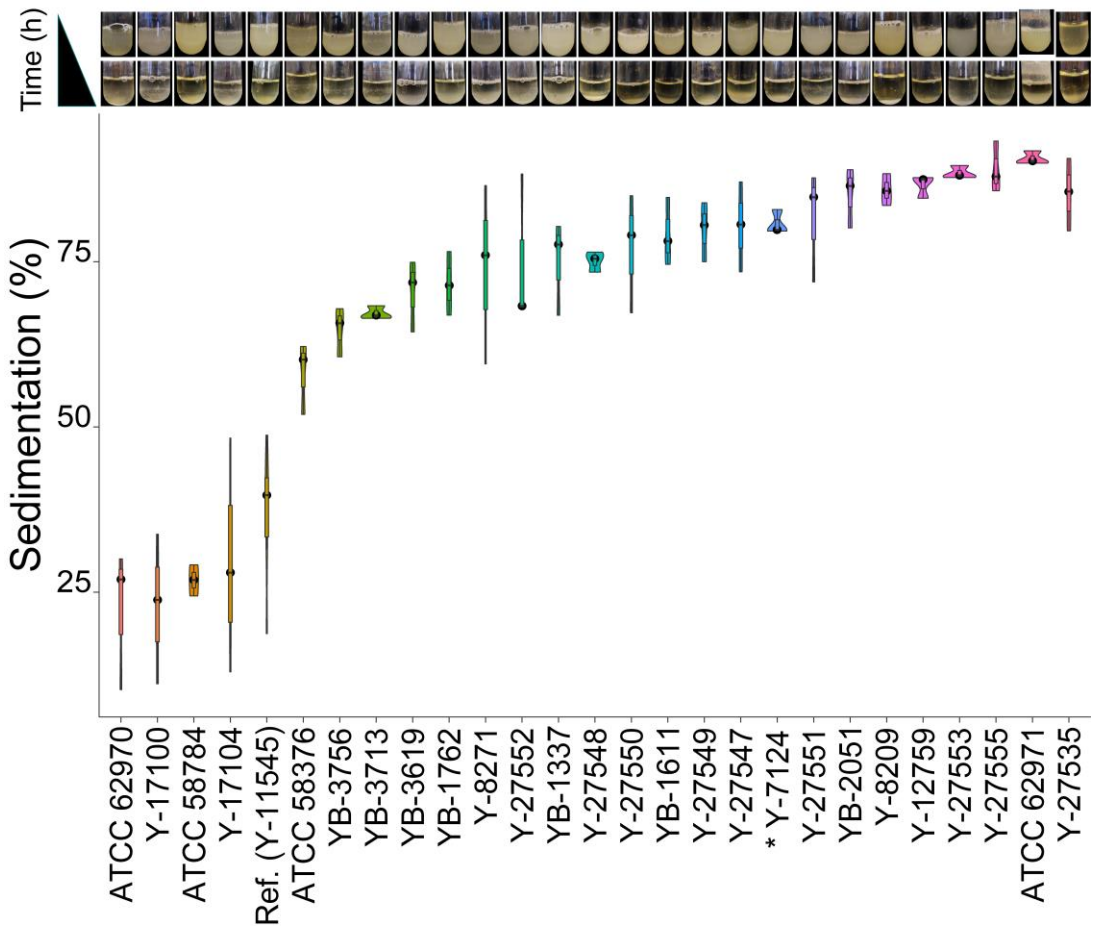


Figure 4.13. Sedimentation (%) profile of natural isolates of natural isolates of *S. stipitis* in SC-Mix. Experiments were conducted measuring the decrease in OD_{600} in spectrophotometer cuvettes with cultures standardised at initial OD_{600} = 1.0±0.1 for 4 hours. All the strains reached their maximum sedimentation after 4 hours, except Y-27535, which reached it after 2 hours. Experiments were conducted in three biological replicates. Ref: Reference strain in genetic studies (Y-11545). * Indicates the strain Y-7124 sequenced during this study.

The natural isolate of *S. stipitis* that showed best sedimentation properties was Y-27535, since the maximum sedimentation values were reached after two hours of experiment, whereas the rest of the strains were evaluated after four hours. The strain Y-7124 showed improvements in sedimentation respect to the reference Y-11545. The same behaviour was observed for the strains ATCC 62971, Y-27549, YB-1611, Y-12759, Y-27535, Y-8209, Y-27555, YB-2051 and Y-27553, whereas the rest of the strains had statistically the same percentage of sedimentation after four hours (one-way ANOVA $p<0.05$) (**Figure 4.13**).

4.3 DISCUSSION

This project included the study of the genomic karyotype of 27 *S. stipitis* natural isolates and demonstrated that there was not a direct correlation with the habitat from which the strains had been isolated (see chapter 3). However, little is known about how phenotypes in *S. stipitis* are influenced by its genomic organization or by the isolation habitats.

Therefore, the aim of this chapter was to test and understand whether there is a connection between different genome organizations in *S. stipitis*, habitats from which the strains were isolated and phenotypes with potential interest for bioethanol fermentation research.

To do so, the effects of different carbon sources on the growth kinetics (growth rate, maximum OD_{600} and lag time) of all the natural isolates were analysed. Moreover, two biofilm formation related traits (agar invasion and sedimentation) were assessed.

Consequently, the strains were grouped according to their habitats and the variability of phenotypes between and within each group was compared for each parameter (**Figure 4.14**). A strong correlation between phenotype and habitat would be supported by low variability within each group but variability between them.

No clear correlation pattern is observed for the strains according to their habitat, since the groups that show less internal variation are formed by only one strain (Y-17104 for fermentation reactor, Y-12759 for soil and Y-11545 for unknown), whereas the groups with more strains show higher variability. Several studies in yeasts (such as *S. cerevisiae, C. albicans,* or *S. pombe*) have demonstrated that some phenotypic traits are associated to the original habitat of the strains assessed (Warringer *et al.* 2011; Yuan O. Zhu, Sherlock and Petrov 2016; Gallone *et al.* 2016) (Lan *et al.* 2002; Mandelblat *et al.* 2017; Cavalieri *et al.* 2018; Hirakawa *et al.* 2015) (Brown *et al.* 2011; Jeffares *et al.* 2015). However, this homogenization in the phenotypes might be associated with common evolutionary history and selective pressure, and therefore dependant on their genomic variability.

Therefore, since no evident relation was identified between the phenotypes studied and the habitat from which the strains had been isolated, the classification was conducted according to the genome organization observed by CHEF electrophoresis (**Figure 4.15**). This will help to elucidate if similar genome structures condition phenotype, probably by conditioning gene expression.

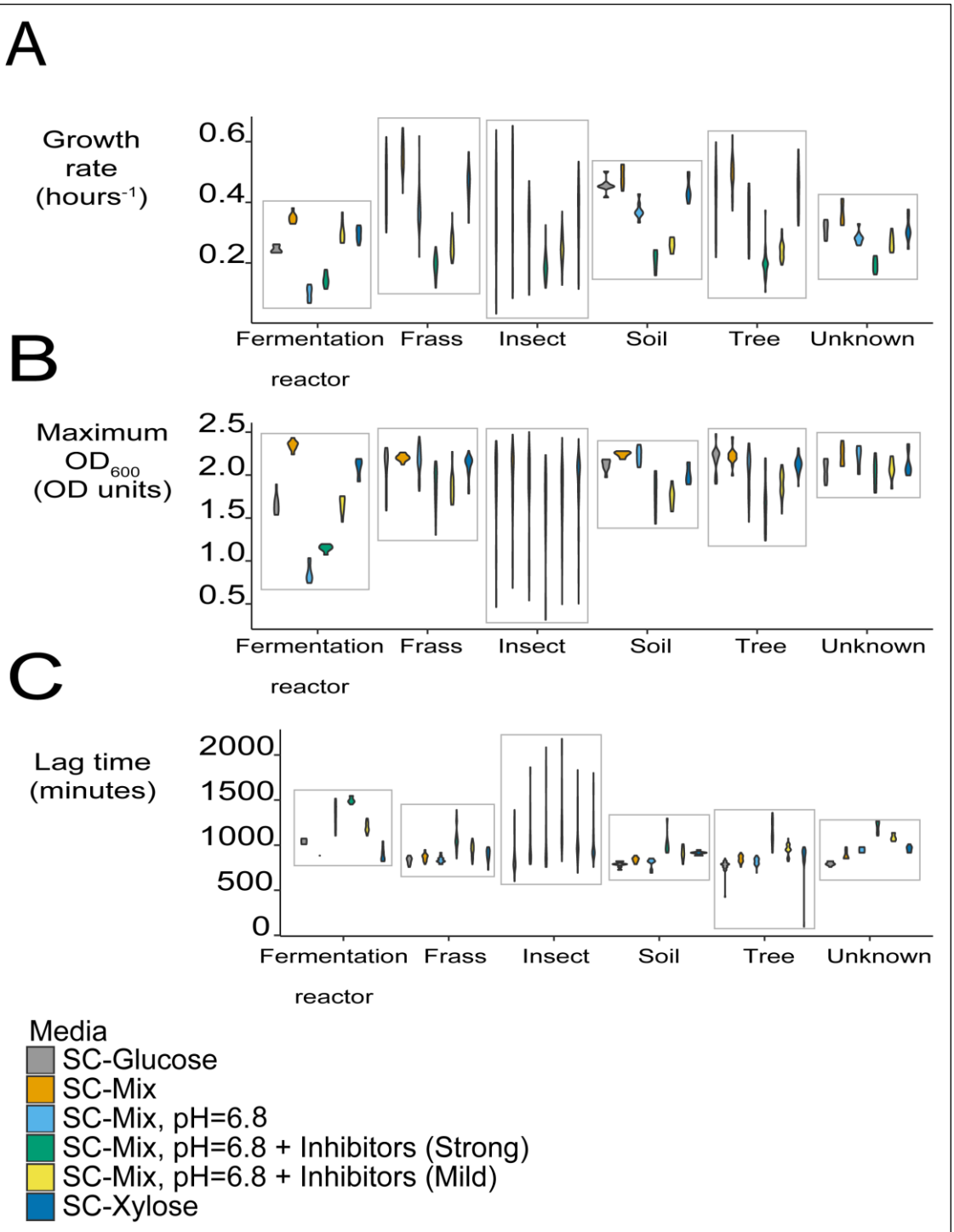


Figure 4.14. Growth rate (hours^{-1}) (A), maximum OD_{600} (OD units) (B) and lag time (minutes) (C) observed for *S. stipitis* natural isolates classified according to their habitats (xylose fermentation reactor, frass, insect, soil, tree or unknown) and for all the different media studied. Experiments were conducted in three technical and three biological replicates.

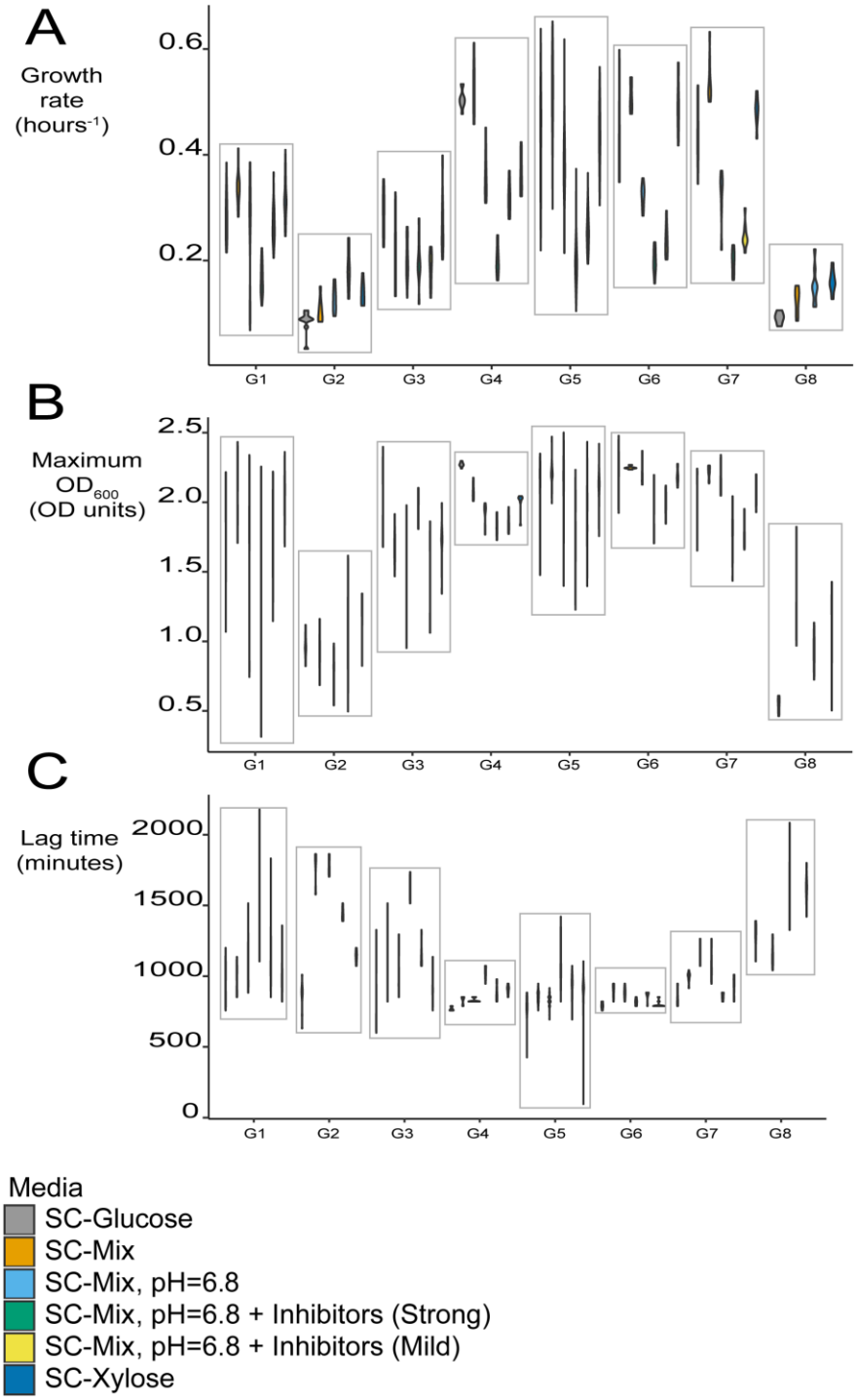


Figure 4.15. Growth rate (hours^{-1}) (A), maximum OD_{600} (OD units) (B) and lag time (minutes) (C) observed for *S. stipitis* natural isolates classified according to their karyotype (CHEF groups G1-G8) and for all the different media studied. Experiments were conducted in three technical and three biological replicates.

In this case, obvious differences were spotted between different groups, which might indicate a correlation between phenotypes and genome organization. Strains with similar genomic organization (belonging to the groups G4, G5, G6 and G7)

showed higher growth rates and lower lag times across all the media studied, whereas strains belonging to G2 and G8 showed slower rates and higher lag times. Although maximum OD_{600} was the least variable parameter, groups G2 and G8 showed again worse results compared to the other groups (**Figure 4.15**).

To understand if the genome structure is related to differences in phenotypes a deeper understanding in how the genomes are organised is required. Prior to this study, only one natural isolate of *S. stipitis* (Y-11545) had been fully sequenced and assembled to chromosome level (Jeffries *et al.* 2007), and consequently, it has been used as a reference strain for this study. Therefore, to understand if any of the genome structures observed show improvements compared to the reference, the growth parameters observed for all the media were normalised to the values obtained for the strain Y-11545 (**Figure 4.16**). Overall, strains that belong the groups G4, G5, G6 and G7 exhibit better growth parameters compared to the reference strain Y-11545, whereas G2 and G8 act notably worse in all the media studied. Although there are not big differences in maximum productivity for natural isolates respect to the reference, the growth rates and lag time are significantly improved for the majority of the strains in G4, G5, G6 and G7, highlighting better growth using both glucose and xylose and the mixture of both as carbon sources.

This finding seems remarkably interesting, since the most used the strains for *S. stipitis* genetic and fermentation studies belong to the groups G1 and G3 (Y-11545 and Y-7124, respectively), which do not offer the best performing parameters assessed in this study.

In the last years, biofilm reactors have gathered attention for their capacity to both enhance the production of value added compounds and the resistance to stress factors commonly found in industrial fermentations (A Demirci, K-LG Ho and AL Pometto III 1997), (Cheng, Demirci and Catchmark 2010).

Some natural isolates of *S. stipitis* tend to form pseudomycelia, and this trait can be improved. Grootjen *et al*. (Grootjen *et al.* 1991) selected strains of *S. stipitis* with enhanced flocculation with the idea of getting higher levels of biomass to process both glucose and xylose in a gas-loop reactor. Moreover, other studies have been conducted with improved flocculent strains of *S. stipitis* (J. P. Delgenes, Moletta and Navarro 1988; De Castro, Oliveira and Furlan 2003). Therefore, the selection of natural isolates that exhibit biofilm formation phenotypes was the next trait had into account in this study.

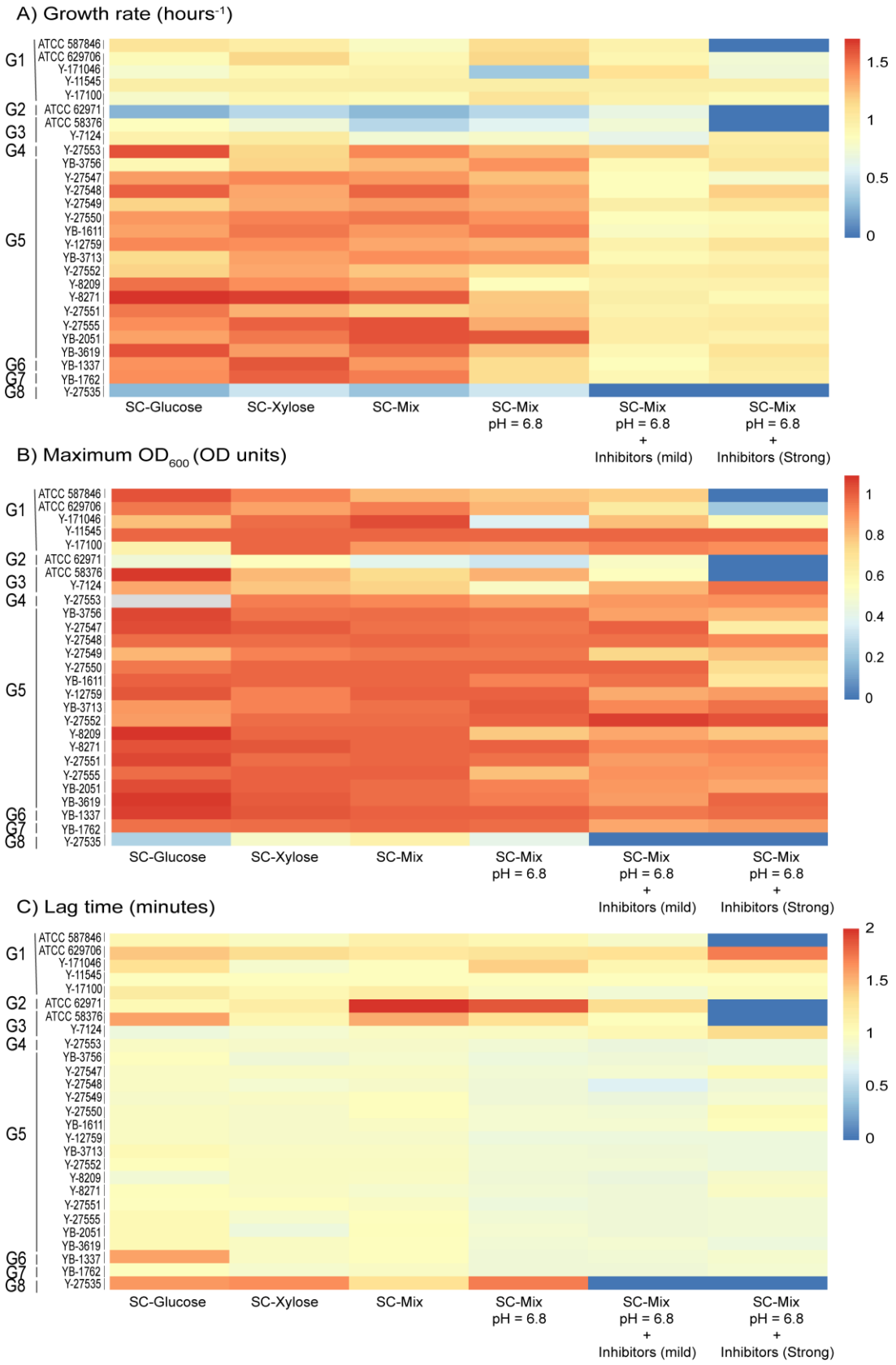


Figure 4.16. Ratio between the average growth rate (hours^{-1}) (A), maximum OD_{600} (OD units) (B) and lag time (minutes) (C) for each *S. stipitis* natural isolate and the reference strain Y-11545 for each media studied. Values higher than 1 indicate increase in the parameter compared to the reference strain, whereas lower indicate a decrease. Natural isolates were classified according to their genome organization (G1-G8).

Similarly, to the results obtained for growth parameters, the classification of the strains according to their habitats did not show a clear pattern of behaviour. However, isolates with different genomic organization exhibit distinct sedimentation profiles, but no variations in agar invasion are detected (**Figures 4.17 and 4.18**).

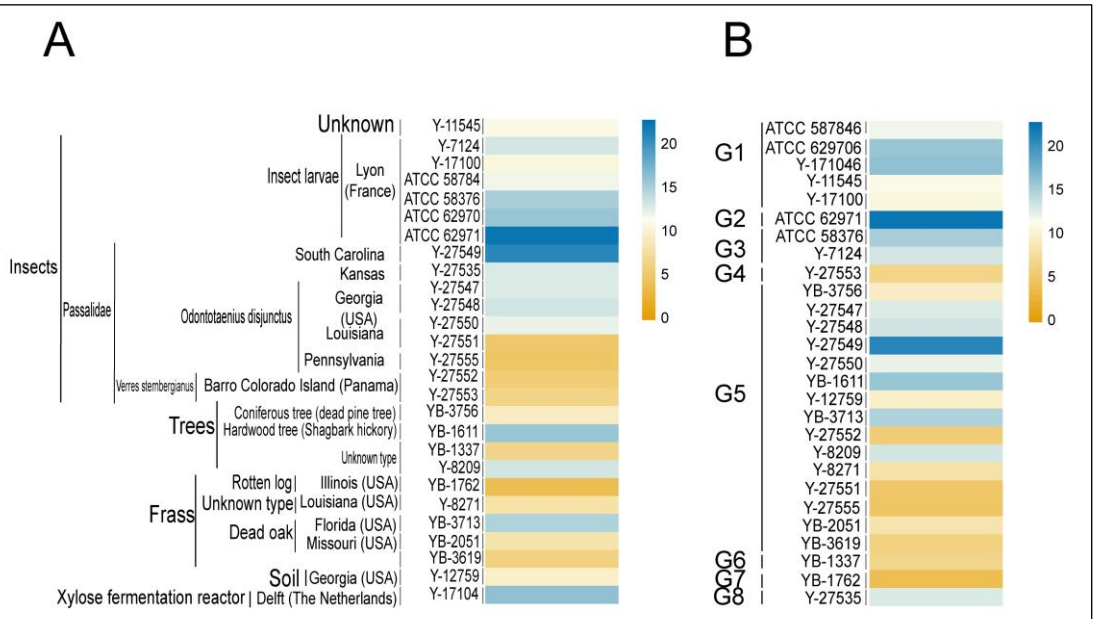


Figure 4.17. Relative agar invasion (%) profiles for each *S. stipitis* natural isolate classified according to their habitat (A) or genome organization (B). No clear pattern is observed for none of the classifications.

Natural isolates obtained from insects seemed to show more agar invasion compared to the other habitats. Also, higher values are observed in genome groups G1 and G2. Despite of this, natural isolates of other groups and habitats also show high agar invasion. Conversely, sedimentation profiles were high for all the genome groups, except G1. The strains present in these groups were isolated from different habitats, which supports the hypothesis of non-habitat related phenotype.

Moreover, the results show that many strains differ in the strength of both traits. Some strains show high percentage of both sedimentation and relative invasion, such as ATCC 62971, Y-27549 and YB-1611, however, others show opposite effects.

To understand if the two phenotypes are related and can provide information about the biofilm related behaviour of each strain, a correlation study was conducted (**Figure 4.19**).

No correlation was observed between the two phenotypes, suggesting that the phenotypic strength of a strain in one assay is not necessarily predictive of its strength across another.

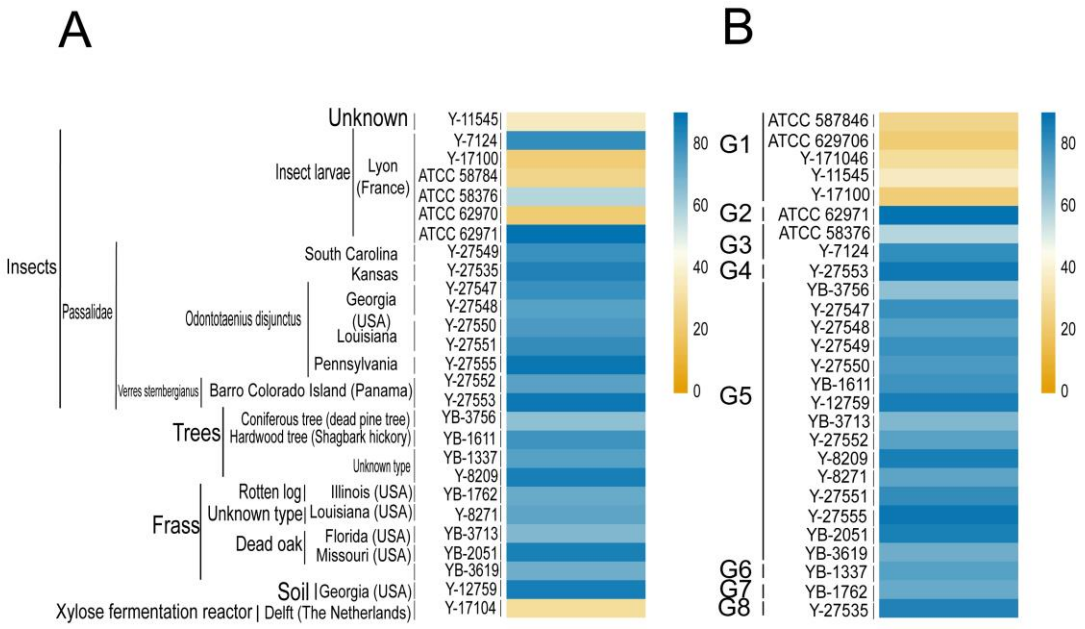


Figure 4.18. Sedimentation (%) profiles for each *S. stipitis* natural isolate classified according to their habitat (A) or genome organization (B). Natural isolates obtained from tress, frass and soil offer high sedimentation percentage, whereas in the case of strains isolated from insects show bimodality (high values (blue) for the isolated from adults and low (brown) for the isolated from larvae) (A). According to the genome organization, strains belonging to G2-G8 offer similar high sedimentation (blue), G1 offers low sedimentation percentage (brown)(B).

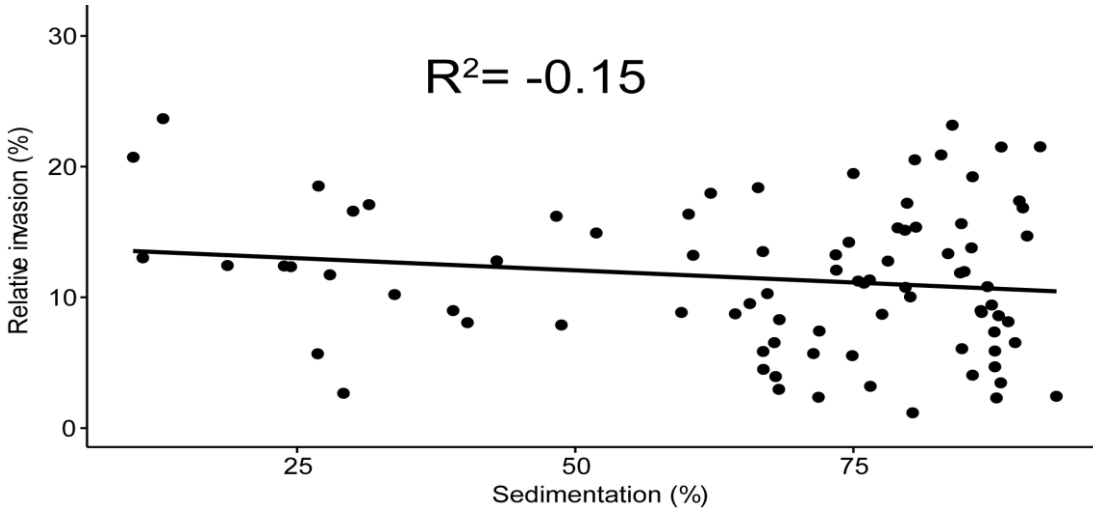


Figure 4.19. Correlation between biofilm related phenotypes in *S. stipitis* natural isolates. Correlation coefficient (R^2= -0.15) indicates no relation between relative invasion (%) and sedimentation (%) profiles (Pearson correlation test).

Despite the lack of correlation between the biofilm related phenotypes studied, the natural isolate ATCC 62971 showed a strong biofilms formation phenotype, with high sedimentation and agar invasion properties. Conversely natural isolates Y-27555 and the reference strain Y-11545 showed weak biofilm formation related phenotypes. This

lack of correlation in biofilm related phenotypes has also been previously described in natural isolates of *S. cerevisiae* (Hope and Dunham 2014).

The strain Y-7124 was sequenced during this study and we have demonstrated that it exhibits a different genome structure when compared to the reference strain Y-11545 (G1 and G3, respectively). Therefore, the analysis of their differences in gene content might help to elucidate variations observed between phenotypic traits in different isolates.

Although their behaviour was similar in general, the strain Y-7124 showed slower growth rate in SC-Mix, lower productivity in SC-Glucose and SC-Mix, and shorter lag time in SC-Glucose. Moreover, it presented higher growth rate under inhibitory conditions, which might indicate better resistance, although the lag time was longer. These differences might be related with non-sense SNPs detected in the strain Y-7124 (**Table 3.3**), affecting sugar transporters (*HXT2.4*), glycosyl hydrolases (involved in the carbohydrate hydrolysis (Davies and Henrissat 1995)).

Finally, the strain Y-7124 showed higher sedimentation percentage compared to Y-11545. Several non-sense SNPs were detected in genes related to hyphal formation (*HWP1*), cell wall synthesis (*KRE9* and *SPF1.3*) and cell wall proteins (*HYR5.5*). Moreover, the duplicated genes present in the deleted region of chromosome 7 in Y-11545, and therefore absent in Y-7124, (*INPT2 and IPT1*) codify for inositol-phosphotransferases, which are essential for inositol metabolism (M. V. Rodrigues *et al.* 2007). Henry *et al.* (Henry *et al.* 1977) have previously demonstrated that inositol starved *S. cerevisiae* cells suffered an arrest in the cell cycle, but kept their metabolism active, which results in an osmotic imbalance that increases the cell density, causing the cell to sediment more rapidly in a ludox gradient. Conversely, *spf1Δ/Δ* mutant strains in *C. albicans* exhibit cell wall defects and a decrease in flocculation (Yu *et al.* 2013), which seems opposite to the effect observed in this study.

Finally, to determine which natural isolate was the best overall performer they were ranked according to the individual phenotypes studied in this section, and a value of 1 (best performer) to 27 (worst performer) was given according to their poisition. The best performer was determined by adding up the scores obtained for each individual phenotype. Therefore, the strains with the lower scores will offer better overall performance (**Table 4.1**).

Table 4.1. Ranking of natural isolates according to the phenotypes studied. G: Karyotype group.

Rank.	Natural Isolate	Score	G	Rank.	Natural Isolate	Score	G
1	Y-27548	137	G5	15	Y-27552	217	G5
2	Y-8209	151	G5	16	YB-3756	219	G5
3	Y-27553	152	G4	17	Y-27549	226	G5
4	Y-27555	159	G5	18	YB-3713	228	G5
5	Y-12759	164	G5	19	Y-11545	236	G1
6	Y-8271	164	G5	20	Y-7124	276	G3
7	YB-2051	174	G5	21	Y-17104	296	G1
8	YB-1337	176	G6	22	Y-17100	322	G1
9	YB-3619	194	G5	23	ATCC 58784	343	G1
10	YB-1762	199	G7	24	ATCC 62970	361	G1
11	YB-1611	200	G5	25	ATCC 58376	375	G3
12	Y-27551	208	G5	26	ATCC 62971	387	G2
13	Y-27550	212	G5	27	Y-27535	405	G8
14	Y-27547	216	G5				

Consequently, the natural isolates Y-27548, Y-8209, Y-27553, Y-27555 and Y-12759 were the top 5 strains of this study. Y-27553 belongs to the genomic group G4 and was isolated from an insect, whereas Y-27548, Y-8209, Y-27555 and Y-12759 belong to the genomic group G5 and were isolated from an insect, tree, insect and soil respectively. Therefore, strains belonging to the genomic group G5 seem to offer the best traits among the studied.

Conversely, the strains ATCC 58784, ATCC 62970, ATCC 58376, ATCC 62971 and Y-27535 were the bottom 5 strains. ATCC 58784 and ATCC 62970 belong to the genomic group G1, ATCC 62971 to G2, ATCC 58376 to G3, and all of them were isolated from insect larvae. Y-27535 belongs to the group G8 and was isolated from an adult insect.

This seems to indicate that strains related to insect gastrointestinal tract are not always the best strains for all the characterization experiments as might indicate the fact that most strains are isolated from that habitat.

4.4 CONCLUSIONS

The possible relationship between the isolation habitats, genome structure and different phenotypes with potential interest for ethanol fermentation and for biofilm formation was analysed for 27 natural isolates of the yeast *S. stipitis*. To do so, the growth

parameters of all the natural isolates were studied in synthetic media with different sugars as carbon source: glucose (SC-Glucose), xylose (SC-Xylose), or a mixture of both (SC-Mix). Moreover, the effects of inhibitors commonly found in lignocellulose hydrolysates on the growth parameters was studied. Finally, the sedimentation and agar invasion profiles of all the natural isolates were determined, as an indication of their potential to form biofilms.

Previous results of this study had failed to demonstrate a correlation between the isolation habitat of the strains and their genomic conformation (See chapter 3). This lack of relationship is also observed between isolation habitat and the phenotypic traits studied in this section. Conversely, phenotypic traits were clustered according to their karyotypic organization observed by CHEF electrophoresis. Strains belonging to groups G4, G5, G6 and G7 offered overall better growth parameters. Only strains belonging to the karyotype organization G1 showed low sedimentation percentage, whereas G2 was the highest. No clear distribution was observed for agar invasion for none of the classifications studied (habitat or karyotype).

Natural isolates Y-27548, Y-8209, Y-27553, Y-27555 and Y-12759 offered the overall best results, with four of those strains (Y-27548, Y-8209, Y-27555 and Y-12759) belonging to the genome organization G5. This is, therefore, an indication that natural isolates with this genome profile could be more suitable for scale up processes.

4.5 FUTURE WORK

The phenotypes studied during this section offer an overall view on how the different natural isolates behave in the presence of different carbon sources and how common inhibitors affect them. However, for a simplification process, the media studied was synthetic. Studying the growth in lignocellulose hydrolysates, scale-up experiments, and specially, the study of the fermentation profile would add value to the screening of natural isolates with potential for industrial bioethanol production.

Finally, to affirm a direct relation among the different genome structures identified in the natural isolates and the differences observed in phenotypes their whole genome sequencing would be required. This would offer: (i) More resolution in the classification of strains in groups, since the low resolution offered by CHEF electrophoresis would be overcome and new differences could be spotted. (ii) Direct information on gene content SNPs and genome modifications that affect phenotypes of interest. For this purpose, information on genome expression levels could also be important and explanatory.

Chapter 5

Genomic and phenotypic modifications in *in vitro* evolved isolates of *Scheffersomyces stipitis* NRRL Y-7124

5.1 INTRODUCTION

Previous results of this study have demonstrated that the genome of *S. stipitis* is highly plastic and that changes in repeats number as well as structural variations are responsible for pervasive genome diversity over long evolutionary time scale. Moreover, we have also demonstrated that several phenotypic traits with potential interest for bioethanol production are conditioned by the genome structure observed in the strains.

Nevertheless, it is still unknown whether *S. stipitis* genome plasticity, and its associated genome diversity, leads to rapid environmental adaption and to increased fitness.

Adaptive laboratory evolution (ALE) has been widely used to analyse the force of evolution in controlled laboratory conditions, since they allow a clear association between phenotype changes and the particular environment that led the trait selection (Dragosits and Mattanovich 2013).

Different adaptation procedures have been previously used for the isolation *S. stipitis* strains with improved performance in different hydrolysates (Nigam 2001b; Nigam 2001a; Bajwa *et al.* 2009; Bajwa, Pinel, Vincent J. J. Martin, *et al.* 2010; Hughes *et al.* 2012). However, to our knowledge, the only previous re-sequencing study on *S. stipitis* strains with superior fermentation phenotypes after adaptation failed to identify the genetic determinants responsible for these improvements. Only 14 point mutations were detected (of which 1 was in a gene related to ethanol fermentation, *ALD7*) but their relation with the phenotype improvement was not further studied (Douglas R. Smith *et al.* 2008).

Slininger *et al.* isolated a total 33 strains derived from the adaptation of the natural isolate *S. stipitis* NRRL Y-7124 to two different industrially promising medias: ammonia fibre expansion-pretreated corn stover hydrolysate (AFEX CSH) and enzymatically saccharified dilute acid pre-treated post-frost switchgrass (SGH), whose solids were separated from the liquor by centrifugation or filtration to produce dilute acid-pretreated switchgrass hydrolysate liquor (PSGHL) (Slininger *et al.* 2015).

This adaptation was conducted by exposition of the strain Y-7124 to increasing concentrations of the medias in a microplate. First, a dilution series of 12% glucan AFEX CSH was used for 122 days, from which the strain Y-50871 was isolated. Then this strain was further challenged to high concentrations of ethanol in an optimum defined media (ODM) with xylose as the only carbon source for 205 days, which resulted in the isolation of the strains Y-50861 and Y-50862. After that, all the strains isolated from this adaptation, and the parental were subjected to enrichment in PSGHL, but only the strain Y-50859, derived from the parental strain after 486 days, showed enhanced fermentation phenotypes, indicating that strain adaptation to one media does not correlate with

general improvements (**Figure 5.1**). During this adaptation improved fermentative phenotypes were observed in the isolates, but the genetic determinants underlying these changes remain unknown.

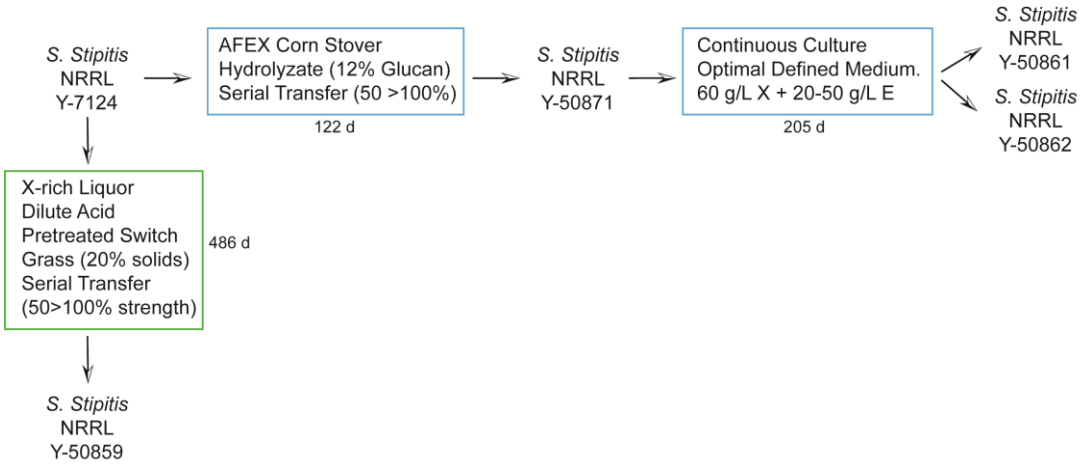


Figure 5.1. *Scheffersomyces stipitis* adaptation flow chart for the isolation of the strains studied in this project (time is measured in days). Abbreviations: AFEX: Ammonia fibre expansion-pretreated. Adapted from Slininger *et al.* (2015)

The aim of this section was to study whether genome modifications are observed in the strains that showed the best fermentation phenotype improvements and, if so, to determine to which extent these modifications are due to genome plasticity and repetitive DNA. To this end, the genome of 4 *S. stipitis* isolates (Y-50859, Y-50861, Y-50862 and Y-50871) generated by exposing the parental strain NRRL Y-7124 to industrially promising medias was analysed by karyotyping, and strains with modifications were further selected for whole genome sequencing.

5.2 RESULTS

5.2.1 Genome structural variations following a real time evolution experiment

To assess whether genome modifications are observed during *in vitro* strain adaptation, the karyotype of *S. stipitis* isolates obtained after adaptation to two industrially relevant media was analysed.

Although the species identification of the evolved isolates was conducted by Slininger *et al*. by sequencing the D1/D2 domain before phenotyping (Slininger *et al.* 2015), the first step in this section was the confirmation of the species identity using the regions spanning the internal transcribed spacers ITS1 and ITS2 and the 5.8S gene (5.8S-ITS rDNA) (**Table 5.1**). The accuracy of this approach was previously confirmed

by Villa-Carvajal *et al.* and previously in this project (See 3.2.1) (Villa-Carvajal, Querol and Belloch 2006).

Table 5.1. Identification of *S. stipitis* evolved isolates by Sanger sequencing of the 5.8S-ITS rRNA gene.

Code	Region	Blast result	% identity	E-value
NRRL Y-7124		*Scheffersomyces stipitis*	99.66	0.0
NRRL Y-50871	5.8S-ITS	*S. stipitis*	99.66	0.0
NRRL Y-50859	rRNA gene	*S. stipitis*	99.66	0.0
NRRL Y-50862		*S. stipitis*	100	0.0
NRRL Y-50861		*S. stipitis*	99.83	0.0

Subsequently, to test whether evolution is linked to large genomic modifications, the karyotype of all the isolates was analysed by CHEF electrophoresis. (**Figure 5.2**).

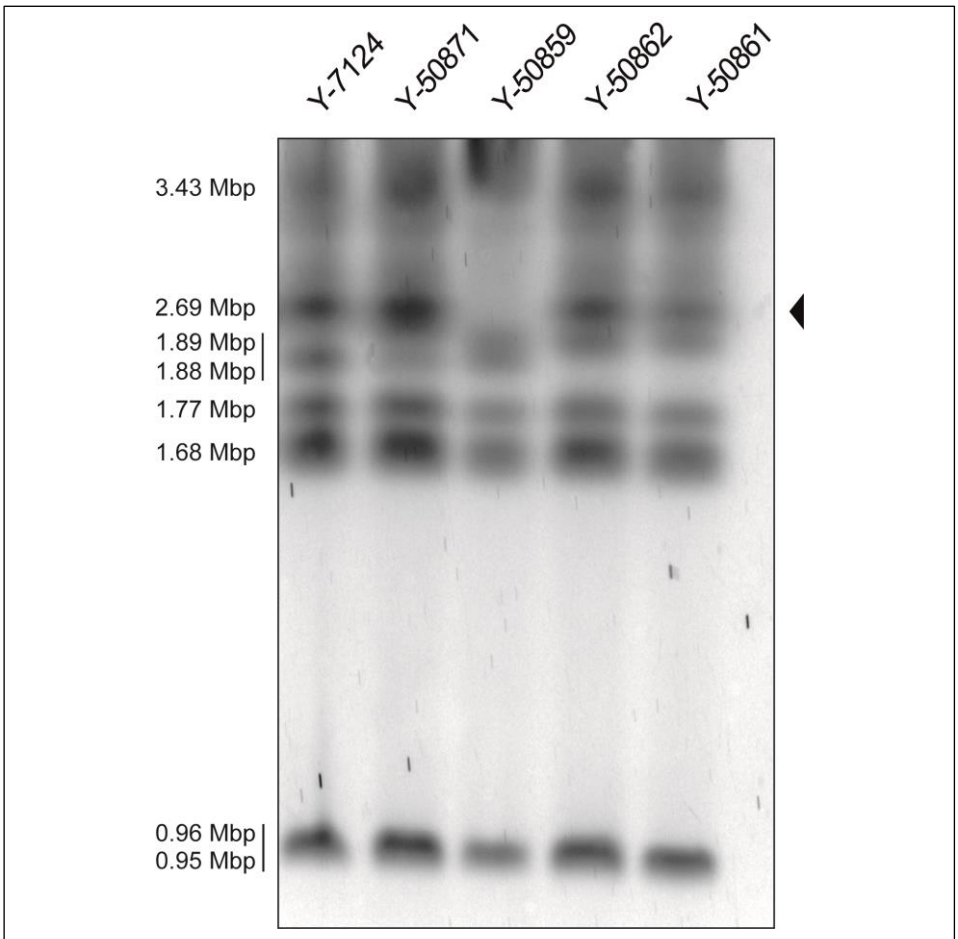


Figure 5.2. CHEF electrophoresis of strains of *S. stipitis* evolved from the natural isolate Y-7124. Running conditions: 6 V/cm, switch time: 60-120s for 12 hours. 4.5 V/cm, switch time: 120-300s switch time for 12 hours at 4.5 V/cm. Staining in ethidium bromide (0.5 µg/ml).

The karyotype of the *in vitro* evolved strains Y-50871, Y-50862 and Y-50861 shows no differences with respect to the parental Y-7124. Only the strain Y-50859 presents the absence of one band, with respect to the parental strain. However, the nature of that modification is not known.

Although the modifications observed during *in vitro* adaptation to industrially relevant medias seem less drastic than those observed in this project on laboratory conventional media (See chapter 3), these results still prove that continuous growth in stressful conditions can also lead to genome modifications.

5.2.2 Genome organisation of strains evolved from the parental *S. stipitis* Y-7124 natural isolate

Karyotype analyses by CHEF electrophoresis revealed that adaptation in hostile environments can lead to genome variation, as the Y-50859 evolved strain presents a chromosome organisation that is distinct from the parental strain Y-7124. Given that the resolution of this technique is too low to detect smaller genomic changes, we sought to unequivocally identify the genomic changes occurred during the evolution experiments by sequencing the genome of two evolved isolates: Y-50859 and Y-50861, as a control with no observable genome reorganizations.

The genomes of the evolved strains Y-50859 and Y-50861 were sequenced following a hybrid TGS approach, based on the combination of Illumina and ONT technologies (with a coverage of 145.17x and 158.09x, respectively). The genomes were organised in 11 and 9 contigs respectively, and they were manually assembled by the identification of centromeres to a total of 8 chromosomes. Therefore, there are no variations in the total number of chromosomes with respect to the parental strain Y-7124. The total genome size for Y-50859 is 15.64 Mbp, whereas it is 15.36 Mbp for Y-50861. This means that the genome size of the evolved strains is slightly larger than the parental strain (15.24 Mbp) in both cases (**Figure 5.3, B and Tables S1 and S2**).

Evolution in this hostile environment resulted in the acquisition of very few point mutations. A total of 14 SNPs (of which 3 are missense) have been detected for Y-50859 (1 variant every 395,779 bases), whereas for Y-50861 a total of 30 SNPs are present (1 variant every 480,636), of which 11 are missense and 1 is nonsense (**Table 5.2**).

Interestingly, all 4 genes affected by SNPs in the strain Y-50859 are also affected in Y-50861, presenting in both cases the same modifications (1 nucleotide change leading to a synonymous SNP and 3 changes in protein residues) (**Table 5.3**). Therefore, modifications for both strains were observed in the proteins codified by *OCA1* (a protein tyrosine/serine phosphatase), by *PICST_59269*, a transporter of the major facilitator superfamily, and by *PICST_31552*, a PAP-1 binding protein. Considering all SNPs are

present in both strains, there is a possibility that the mutations might be present in the parental strain Y-7124.

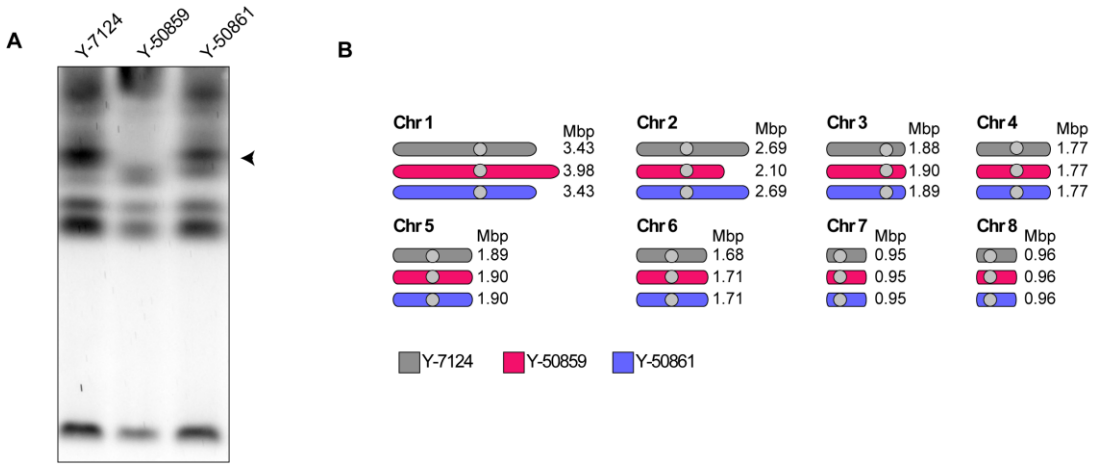


Figure 5.3. CHEF electrophoresis karyotype (A) and chromosome organization and size (B) for parental (Y-7124) and evolved strains (Y-50869 and Y-50861).

Table 5.2. Summary of the SNPs observed between the parental strain *S. stipitis* NRRL Y-7124 and the two *in vitro* evolved strains Y-50859 and Y-58061.

	Y-50859	Y-50861
Number of variants	14	30
Variant rate	1 variant / 395,779 bases	1 variant / 480,636
Transitions	6	10
Transversions	8	20
Region		
Exon	4 (28.57%)	14 (46.67%)
Intergenic	10 (71.43%)	16 (53.33%)
Effects		
Missense	3 (75.00%)	11 (78.57%)
Nonsense	1 (25.00%)	1 (7.14%)
Silent (synonymous)	-	2 (14.29%)

The strain Y-50861 presents 10 additional SNPs (1 synonymous, 8 missense and 1 nonsense). The gene *PICST_56002*, which codifies for a GTP-binding ADP ribosylation factor, presents the accumulation of 4 missense SNPs and the only nonsense mutation reported in the evolved strains. Moreover, the genes *PICST_62836* and *PICST_73970*, which encode for an MCP-domain signal transduction protein and a predicted P-loop ATPase fused to an acetyltransferase respectively, are also affected by missense SNPs. The rest of SNPs affect to genes whose protein product has not been characterised (**Table 5.3**).

Table 5.3. List of genes in *S. stipitis* evolved strains Y-50859 and Y-50861 affected by SNPs

Strain	Gene ID	Function	Effect	Modifications
Y-50859	PICST_46361 (OCA1)	protein tyrosine/serine phosphatase	Missense	c.358C>T|p.His120Tyr
	PICST_59269	MFS transporter	Missense	c.161G>A|p.Gly54Glu
	PICST_31552	PAPA-1 domain containing protein (PAP-1 binding protein)	Missense	c.87C>A|p.Asp29Glu
	PICST_32448	Predicted protein	Synonymous	c.21C>G|p.Ala7Ala
Y-50861	PICST_46361 (OCA1)	protein tyrosine/serine phosphatase	Missense	c.358C>T|p.His120Tyr
	PICST_59269	MFS transporter	Missense	c.161G>A|p.Gly54Glu
	PICST_31552	PAPA-1 domain containing protein (PAP-1 binding protein)	Missense	c.87C>A|p.Asp29Glu
	PICST_56002	GTP-binding ADP-ribosylation factor	Nonsense (1) Missense (4)	- c.199C>A|p.Gln67Lys c.194G>C|p.Gly65Ala c.191G>C|p.Cys64Ser c.176T>A|p.Phe59Tyr
	PICST_49608	Predicted protein	Missense	c.68G>T|p.Arg23Ile
	PICST_62836	MCP-domain signal transduction protein	Missense	c.1442C>A|p.Thr481Asn
	PICST_33084	Predicted protein	Missense	c.1226A>T|p.Asn409Ile
	PICST_73970	Predicted P-loop ATPase fused to an acetyltransferase	Missense	c.237G>T|p.Lys79Asn
	PICST_32448	Predicted protein	Synonymous	c.21C>G|p.Ala7Ala
	Gene_02570 (No transcript ID detected)	-	Synonymous	c.264C>G|p.Thr88Thr

Despite the similarity observed at single polymorphisms, several important genomic changes have occurred in the evolved strains following passaging in challenging growth conditions. Chromosome structure comparisons by Mummer alignments revealed that the evolved Y-50859 strain is characterised by a translocation between chromosome 1 and 2 (**Figure 5.4, A**).

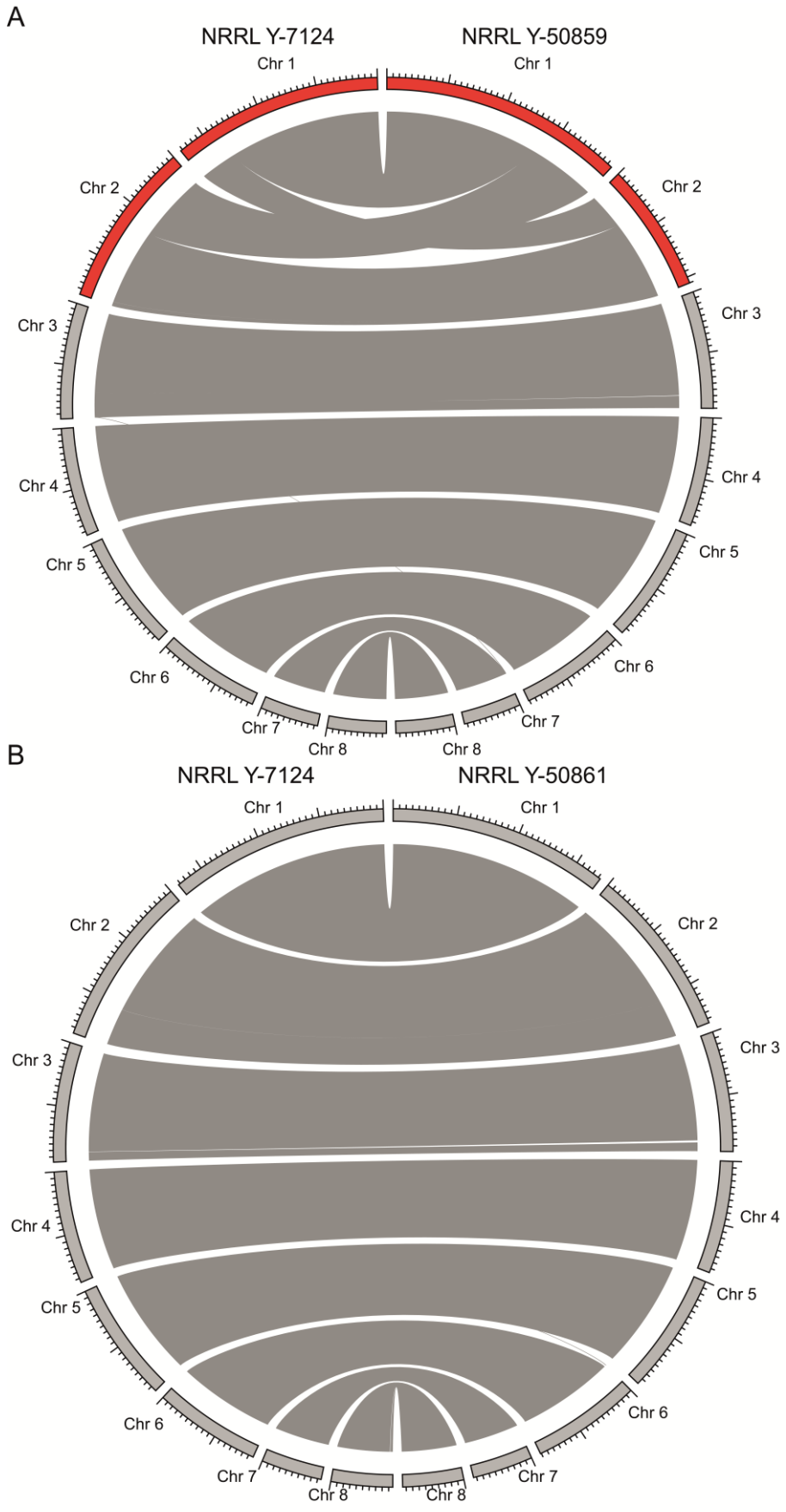


Figure 5.4. Whole genome alignment (Mummer) between the natural isolate *S. stipitis* Y-7124 and the derived strains Y-50859 (A) and Y-50861 (B). Red indicates chromosomes where rearrangements are observed.

PCR analyses with primers spanning the breakpoint confirm this result **(Figure 5.5**). However, in this case, analysis of the translocation breakpoint fails to identify any repetitive element in the location. On the other hand, no conformational changes are observed between the parental and evolved strain Y-50861 (**Figure 5.4, B**).

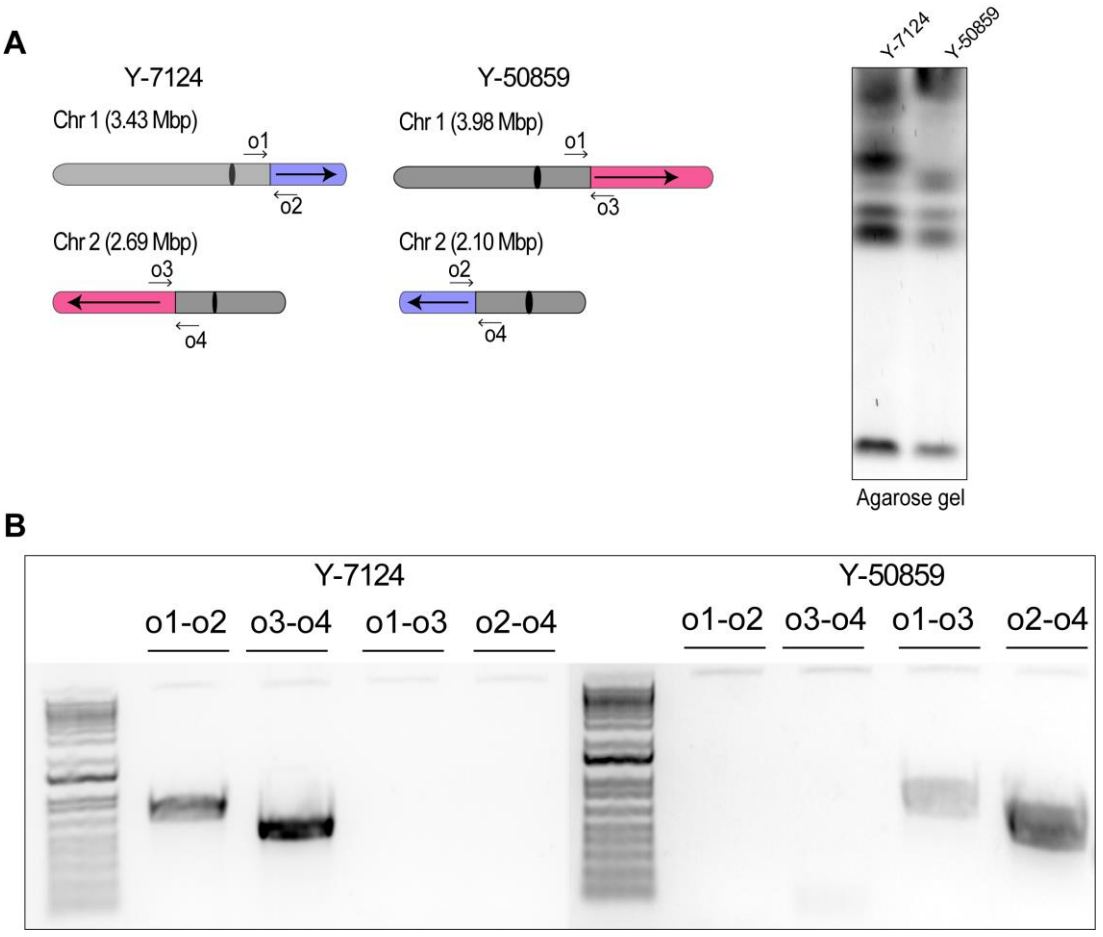


Figure 5.5. Map of the primers used to validate the translocation observed by genome alignments and CHEF electrophoresis between chromosomes 1 and 2 in the evolved strain Y-50859 (A) PCR results spanning the translocation breakpoint in both parental (Y-7124) and evolved strain (Y-50859) (B). o1: AB968, o2: AB969, o3: AB966, o4: AB967.

The translocation observed between chromosome 1 and 2 in the evolved strain Y-50859 leads to truncation of one ORF, XP_001387101, codified by the unannotated gene *PICST_53805*. The presence of this complete gene is not detected in the genome of the strain Y-50859, although the two fragments in which it is divided through the breakpoint are detected in chromosome 1 and 2, respectively.

5.2.3 Genome modifications in repetitive elements during real time evolution

Previous results of this study have demonstrated the variability of repetitive regions during evolution in two natural isolates of *S. stipitis.* Variations in subtelomeres, centromeres and transposable elements were detected and described (See Chapter 3).

However, the effects of repetitive elements in short-term *in vitro* evolution of *S. stipitis* have not been described to date.

No differences were observed in the distribution and location of transposable elements between the three strains, with a 100% conservation in all 20 TEs detected for the parental strain Y-7124 (**Figure 5.6, Tables S5 and S6**). Moreover, no important variations were observed in the number of transposon-associated repeats.

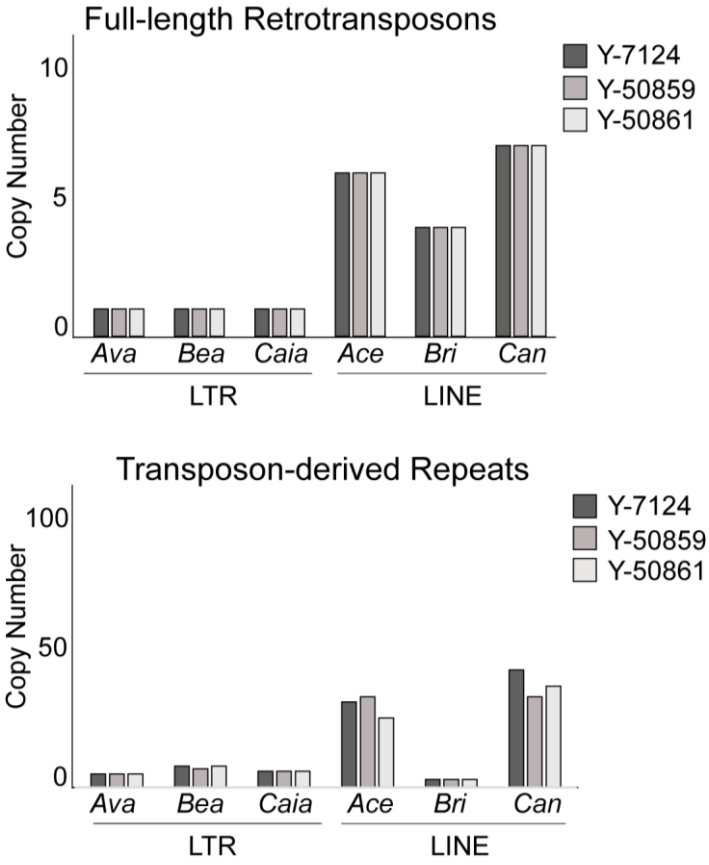


Figure 5.6. Copy number of full-length retrotransposons and transposon-derived repeats in the *S. stipitis* parental strain Y-7124 and evolved strains Y-50859 and Y-50861.

No variations were observed either in the subtelomeric regions of the three strains. The orientations of all the genes are maintained, except for the two chromosome arms involved in the reciprocal translocation between chromosome 1 and chromosome 2 in the strain Y-50859 (**Figure 5.7, A**). Although the number of genes in each family studied has variations, these seem related to assembly errors, since in some cases they have been split into several ORFs whose combination maintains the size of the original complete gene (**Figure 5.7, B**).

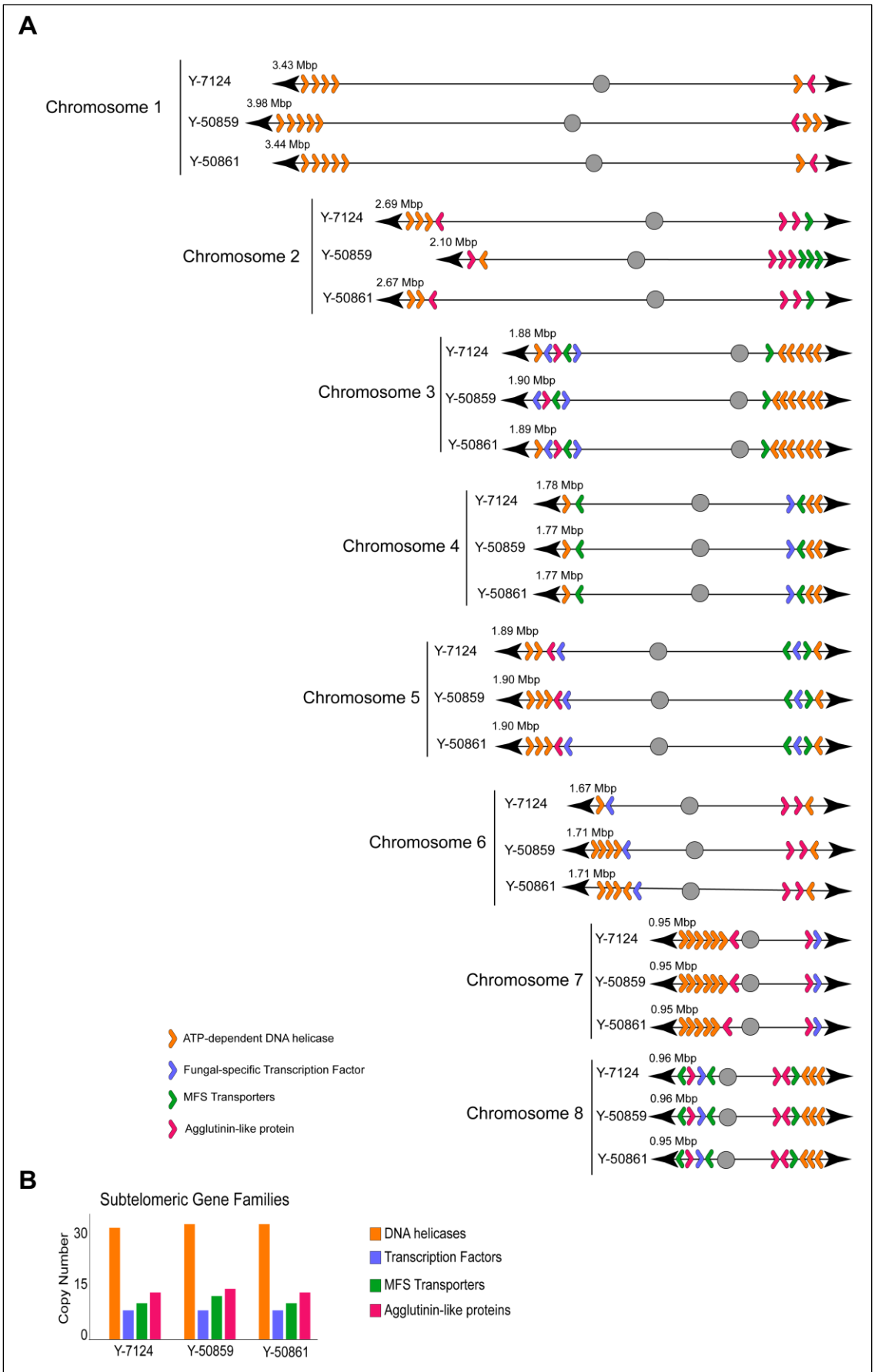


Figure 5.7. Distribution of subtelomeric gene families in *S. stipitis* natural isolate NRRL Y-7124 and evolved strains Y-50859 and Y-50861 (A) Copy number of genes present in each family (B).

Despite of the lack of variability in the subtelomeres and transposable elements in short-term evolution, the size and organisation of 3/8 centromeres (CEN1, CEN5 and CEN7) is different in the two evolved strains compared to the parental Y-7124 (**Figure 5.8, Table 5.4**).

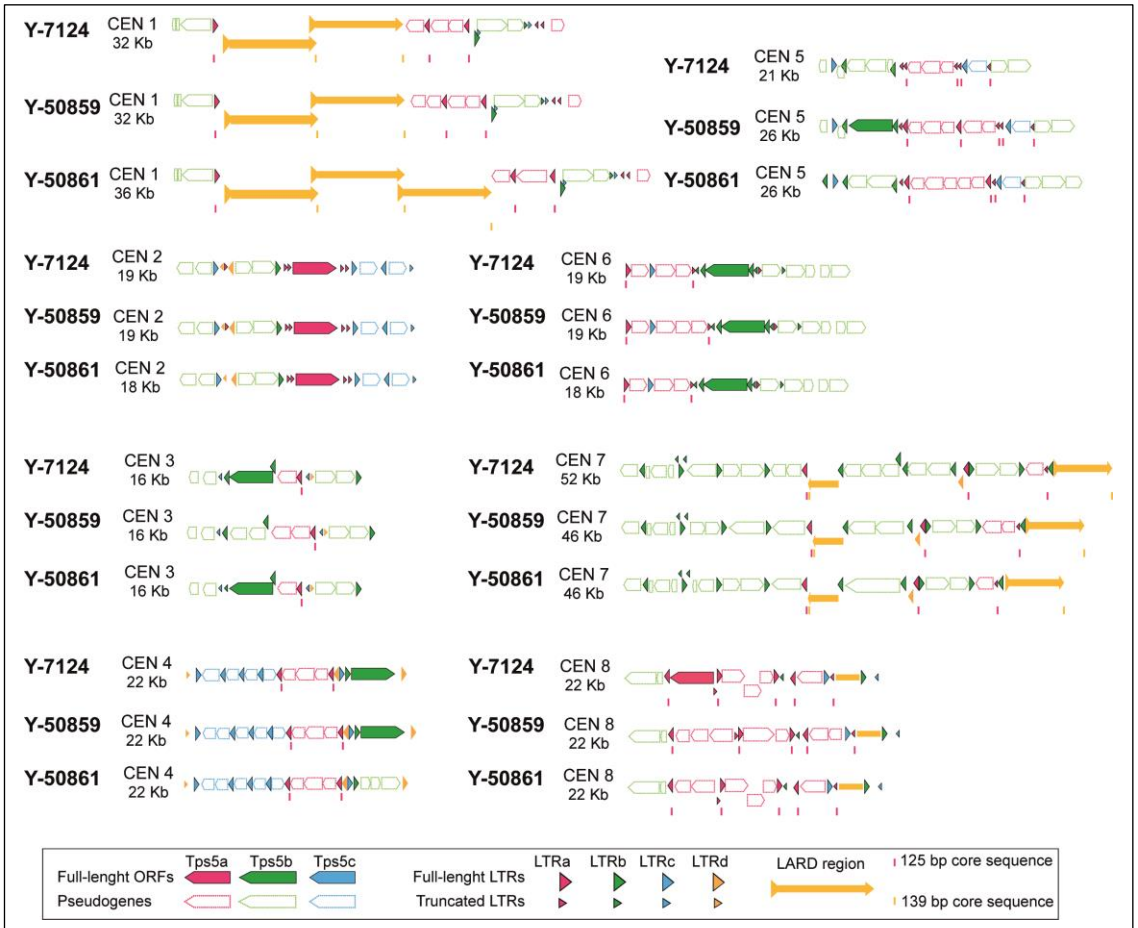


Figure 5.8. Comparison of the centromere structural organization for the 8 chromosomes present in all *S. stipitis* strains studied (parental Y-7124 and evolved strains Y-50859 and Y-50861). Abbreviations: Tps5: Ty-like transposon, LTR: Long terminal repeat, LARD: Large retrotransposon derivative.

Although the average size of the centromeres is maintained for the three strains (25.5±11.8 kb for Y-7124, 25.2±9.7 kb for Y-50859, and 25.6±10.0 kb for Y-50861) (**Table 5.4**), the centromere of chromosome 1 in Y-50861 is approximately 4.0 kb larger, which is related to the presence of an extra LARD adjacent to the two already pre-existent in the parental strain (**Figure 5.8**). Moreover, the centromere of chromosome 5 in both evolved strains is around 5 kb larger, which in both cases seems related to the insertion of a Tps5a transposon in the same position. Finally, the centromere 7 of the parental strain Y-7124 is approximately 7 kb larger, which seems related to the presence of an extra Tps5b. Therefore, the total number of complete Tps5 ORFs and pseudogenes, and the repetitive regions associated to them were slightly different (**Table 5.5**).

Table 5.4. Regional centromere size for each chromosome (C) in the *S. stipitis* strain Y-7124 and its derived strains Y-50859 and Y-50861. Red colour indicates size change with respect to the parental strain.

	NRRL Y-7124	NRRL Y-50859	NRRL Y-50861
C1	31.8	32.2	35.5
C2	18.6	18.6	18.5
C3	16.2	16.2	16.2
C4	22.4	22.4	22.4
C5	21.4	26.2	26.0
C6	18.5	18.5	18.5
C7	52.4	45.7	45.7
C8	22.2	22.3	22.3
average	25.5±11.8	25.2±9.7	25.6±10.0

Table 5.5. Number of complete Tps5 genes, pseudogenes (p), large retrotransposons derivatives (LARD), and complete and truncated (t) long terminal repeats (LTR) distributed in the centromeres of the *S. stipitis* strains Y-7124, Y-50859 and Y-50861 .

	Y-7124	Y-50859	Y-50861		Y-7124	Y-50859	Y-50861
Tps5a	2	1	1	LTRa	13	14	13
Tps5b	3	3	2	LTRb	21	20	20
Tps5c	0	0	0	LTRc	12	12	12
LARD	4	4	5	LTRd	10	10	11
Tps5ap	19	28	22	LTRat	16	14	12
Tps5bp	41	38	40	LTRbt	7	7	7
Tps5cp	7	8	8	LTRct	6	6	6
				LTRdt	3	3	3
Total	76	82	78	Total	88	86	84

Despite of the lack of modification at non-centromeric transposable elements, in their associated repeats and at subtelomeric regions, the changes at the repetitive centromeric regions lead to the conclusion that reorganisation of repetitive elements is also present during short-term evolution.

5.2.4 Analysis of copy number variation (CNV)

To determine whether real time evolution in hostile environments resulted in copy number variation (CNV) of specific genomic regions, the sequence coverage depth of both Illumina and Nanopore reads over the parental strain genome were compared. This read-depth analysis revealed that the evolved Y-50859 strain is characterised by a Chr5 segmental aneuploidy (**Figure 5.9**). Apart from the aneuploid region, several peaks of

coverage were detected along the genome, which are related to centromeric sequences or to positions in which repeats associated to transposons are detected.

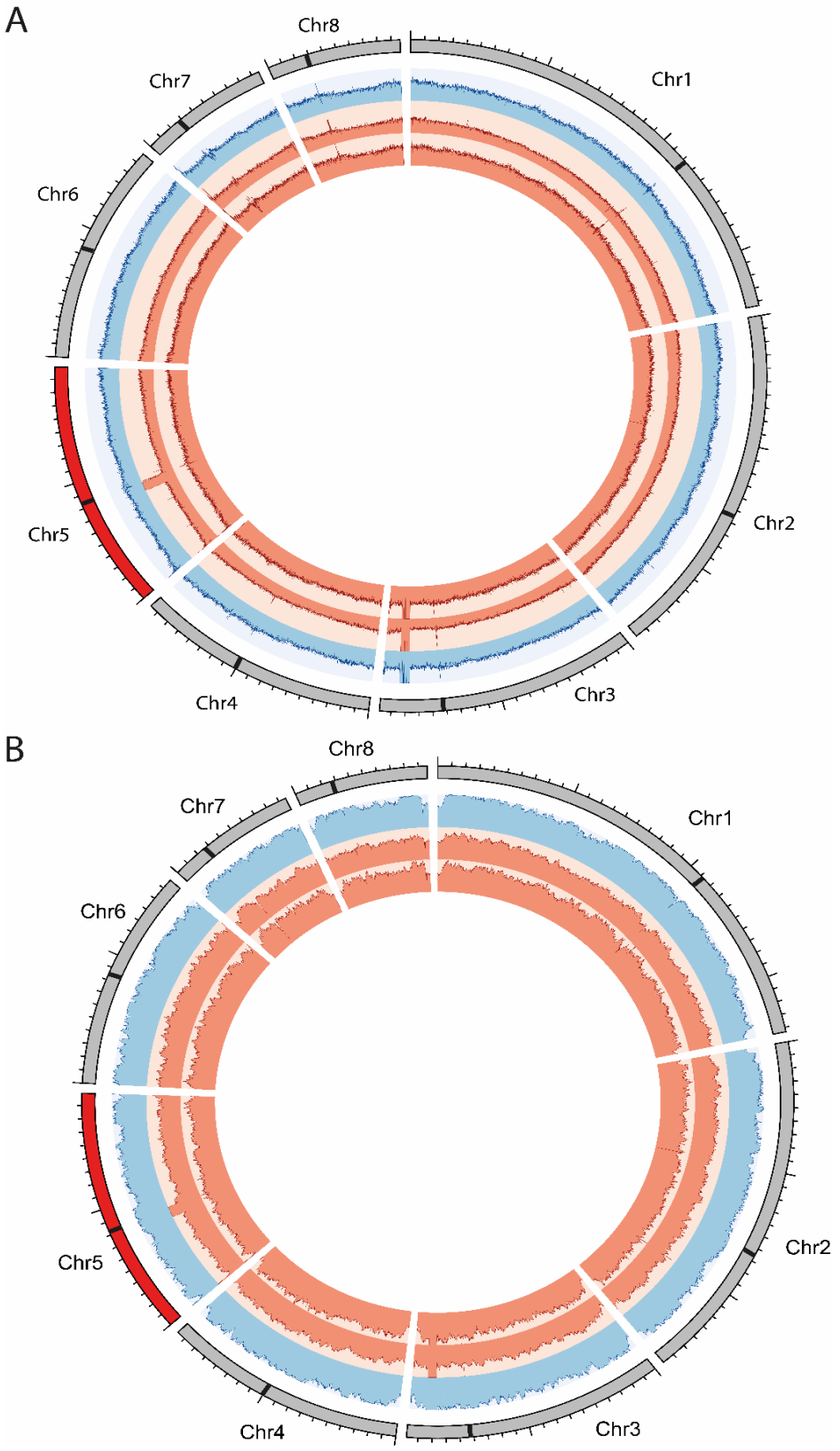


Figure 5.9. Coverage of Illumina (A) and Nanopore (B) reads for the strains Y-7124 (blue), Y-50859 (top red) and Y-50891 (bottom red) over the assembly of the parental strain Y-7124. Red indicates chromosomes with changes in read coverage. Black boxes indicate the position of the centromeres in each chromosome.

Study of the Nanopore sequencing data in the aneuploidy region allows the identification of several distinct features within the segment: High coverage of an approximately 75 kb region of Chromosome 5 (CNV 5L), and high coverage at centromeric sequences of CEN5 (**Figure 5.10**).

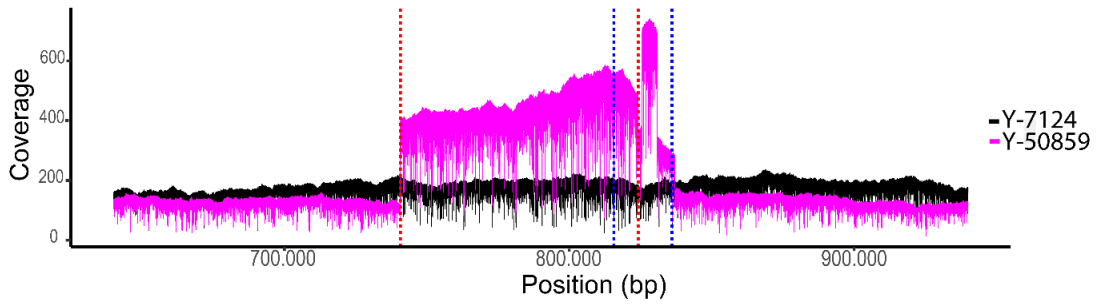


Figure 5.10. Coverage of the Nanopore reads for both parental strain Y-7124 (black) and evolved Y-50859 (magenta) according to the position (bp) where the aneuploidy is detected in chromosome 5 of the parental strain Y-7124. Blue lines indicate the position of the centromere 5, whereas red lines indicate the region of aneuploidy.

Further analysis of the ONT reads revealed the presence of tandem telomeric repeats, and that the Chr5 segmental aneuploidy contains identical DNA sequences, derived from Chr5L, flanking a centromere with different structure to the centromere of chromosome 5 (**Figure 5.11, A and B**). Furthermore, analyses of the Nanopore sequencing data suggested that CNV 5L is not part of the 8 nuclear *S. stipitis* chromosome, as we could not detect any long reads sequence containing the CNV 5L associated with other genomic sequences.

This preliminary information strongly suggested that the observed Chr5 segmental aneuploidy was caused by the presence of a linear isochromosome (i-5L) composed of two identical chromosome arms flanking a centromere. The presence of both centromere and telomeric sequences led us to hypothesise that i-5L is mitotically stable. To test this hypothesis, the CHEF electrophoresis run conditions were adjusted to detect, in addition to the large nuclear chromosome, smaller (20-fold) DNA fragments. This analysis validated our hypothesis as a linear DNA fragment of ~ 180 kb was detected in the evolved strain Y-50859 but not in the parent strain Y-7124 or in the other evolved strain sequenced Y-50859 (**Figure 5.11, C**). Southern analyses confirm that the detected DNA fragment contains sequences with homology to chromosome 5L (**Figure 5.11, C**).

A fluctuation analysis was conducted to determine the stability of the i-5L chromosome. For that purpose, a PCR detection system was developed using primers designed to anneal in the region spanning the telomeric repeats and the isochromosome (**Figure 5.11, A**). Subsequently, 7 colonies of the strain Y-50859 were grown overnight (15 hours) in liquid YPD, plated at a density of 200 cells per plate and grown for 48 hours

at 30 ºC. The presence of the i-5L chromosome was then assessed by PCR in ten colonies of each plate. The experiment was repeated three times, for a total screening of 210 colonies. A PCR for the amplification of actin was used as control. From the total 210 colonies, only 9 (4.29±1.43%) were negative for the presence of the i-5L chromosome. The absence of the isochromosome was further demonstrated by CHEF electrophoresis and Southern blot analysis (**Figure 5.11, C**). The aneuploidy detected might lead to the overexpression of genes present in the region, which could be related with the enhanced fermentation detected in this strain when grown in stressful conditions. However, phenotype improvements might not be present in the absence of stress, since the i-5L chromosome is lost in standard laboratory conditions.

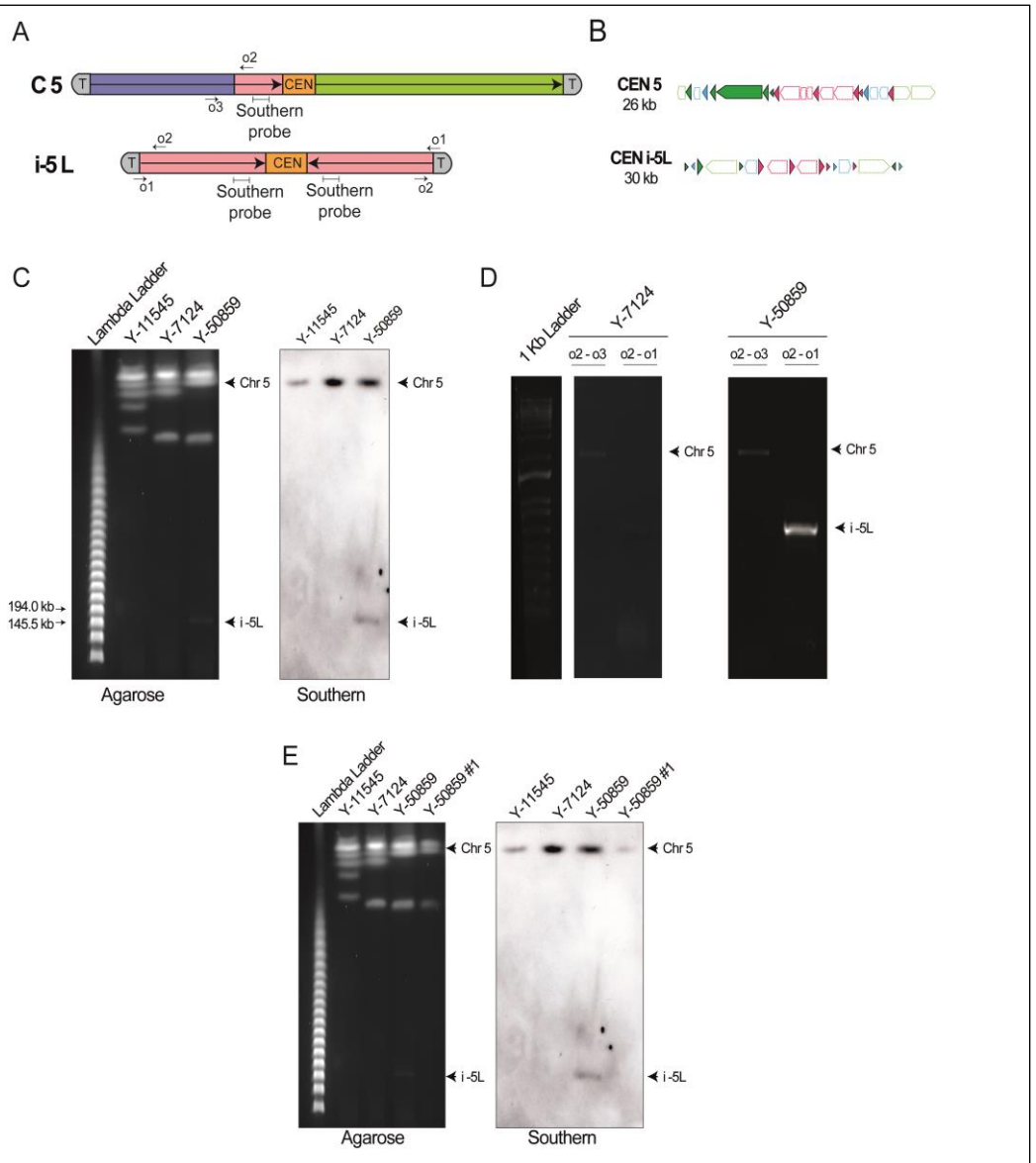


Figure 5.11. Structure of the isochromosome 5L formed by inverted repeats of a region of chromosome 5 in the strain Y-50859 (A) Comparison of the structure of the Centromeres of C5 and i-5L (B) CHEF electrophoresis and Southern Blot analysis showing the presence of i-5L (C) using a probe for C5. PCR system for the detection of the i-5L (D) using the primers mapped in (A). CHEF electrophoresis and Southern Blot showing the absence of the i-5L in a strain derived from Y-50859 (Y-50859#1) (E). o1: AB1019, o2: AB1015, o3: AB1016.

A total of 26 genes were identified as part of the high coverage region (**Table 5.6**). These genes are estimated to be present in 3 copies in the evolved strain Y-50859 (one in the chromosome 5 and two in the isochromosome i-5L). Among these, 8 are involved in metabolic processes, 11 in cellular and signalling processes, and 5 in information storage.

These results demonstrate that propagation in hostile environments can lead to the formation of new chromosomes thanks to the presence of repetitive elements, such as centromeres and telomeres.

Considering that the new chromosome detected is formed of repeated genes, it might act as an important source for genome evolution. Nevertheless, none of the genes present in the i-5L chromosome is essential for growth in non-stressful conditions, since over 4% of the population suffers a chromosome loss in laboratory conditions.

5.2.5 Phenotypic changes associated to evolution

This project has shown that the adaptation of the natural isolate of *S. stipitis* NRRL Y-7124 to two industrially relevant medias, PSGHL and AFEX CSH, can produce modifications in the genome conformation of isolated evolved strains (Y-50859 and Y-50861, respectively). This adaptation also produced improved fermentation phenotypes previously described by Slininger *et al*. (Slininger *et al.* 2015). Considering the difficulty of repeating the fermentation experiments in which the improvements of the evolved strain were observed, the next step was to study possible differences in phenotypes easily measurable with routine methodologies, but that could still be interesting for fermentation industrial research.

To this end, the phenotypic determinations for the natural isolates were repeated for these strains (See chapter 4). These were based on the determination of the growth in different medias by studying the change in optical density over time in microplates (**Figure 5.12**). Subsequently, three growth parameters: growth rate, maximum OD_{600} (as an indication of biomass productivity) and lag time, were calculated for each strain using glucose, xylose, and a mixture (SC-Mix) of glucose (60% of the total sugars) and xylose (40% of the total sugars) as carbon source. Moreover, the effects of growth and fermentation inhibitors commonly present in lignocellulose hydrolysates was also evaluated using two different concentrations of an inhibitor cocktail mixture (mild and strong). Finally, the ability of the strains to form biofilms was also assessed by the determination of both agar invasion and biomass sedimentation.

Table 5.6. Genes identified in the isochromosome 5L (i-5L).

Gene (PICST)	KOG code	Family	Function	Cellular process
22565	KOG3177	Oligoketide cyclase/lipid transport protein	Lipid transport and metabolism	METABOLISM
65754	KOG2616	Pyridoxalphosphate-dependent enzyme/predicted threonine synthase	Amino acid transport and metabolism	
47494	KOG2596	Aminopeptidase I zinc metalloprotease (M18)	Amino acid transport and metabolism	
67821	KOG4476	Gluconate transport-inducing protein	Carbohydrate transport and metabolism	
47531	KOG1380	Heme A farnesyltransferase	Coenzyme transport and metabolism	
78208	KOG2828	Acetyl-CoA hydrolase	Energy production and conversion	
83761	KOG2112	Lysophospholipase	Lipid transport and metabolism	
32069	KOG2499	Beta-N-acetylhexosaminidase	Carbohydrate transport and metabolism	
67817	KOG0715	Molecular chaperone (DnaJ superfamily)	Posttranslational modification, protein turnover, chaperones	CELLULAR PROCESSES AND SIGNALING
67819	KOG2992	Nucleolar GTPase/ATPase p130	Nuclear structure	
67821	KOG4476	Gluconate transport-inducing protein	Signal transduction mechanisms	
46961	KOG1556	26S proteasome regulatory complex, subunit RPN8/PSMD7	Posttranslational modification, protein turnover, chaperones	
90035	KOG2635	Medium subunit of clathrin adaptor complex	Intracellular trafficking, secretion, and vesicular transport	
46396	KOG2223	Uncharacterized conserved protein, contains TBC domain	Signal transduction mechanisms	
83705	KOG4719	Nuclear pore complex protein	Nuclear structure	
83705	KOG4719	Nuclear pore complex protein	Intracellular trafficking, secretion, and vesicular transport	
89662	KOG0185	20S proteasome, regulatory subunit beta type PSMB4/PRE4	Posttranslational modification, protein turnover, chaperones	
32075	KOG0946	ER-Golgi vesicle-tethering protein p115	Intracellular trafficking, secretion, and vesicular transport	
32083	KOG0733	Nuclear AAA ATPase (VCP subfamily)	Posttranslational modification, protein turnover, chaperones	
60417	KOG1767	40S ribosomal protein S25	Translation, ribosomal structure, and biogenesis	INFORMATION STORAGE AND PROCESSING
32082	KOG3152	TBP-binding protein, activator of basal transcription (contains rrm motif)	Transcription	
46807	KOG3613	Dopey and related predicted leucine zipper transcription factors	Transcription	
89595	KOG1009	Chromatin assembly complex 1 subunit B/CAC2 (contains WD40 repeats)	Replication, recombination, and repair	
89595	KOG1009	Chromatin assembly complex 1 subunit B/CAC2 (contains WD40 repeats)	Chromatin structure and dynamics	
59994	KOG2986	Uncharacterized conserved protein	Function unknown	POORLY CHARACTERIZED
46113	KOG2676	Uncharacterized conserved protein	Function unknown	
46396	KOG2223	Uncharacterized conserved protein, contains TBC domain	General function prediction only	
46516	KOG1533	Predicted GTPase	General function prediction only	

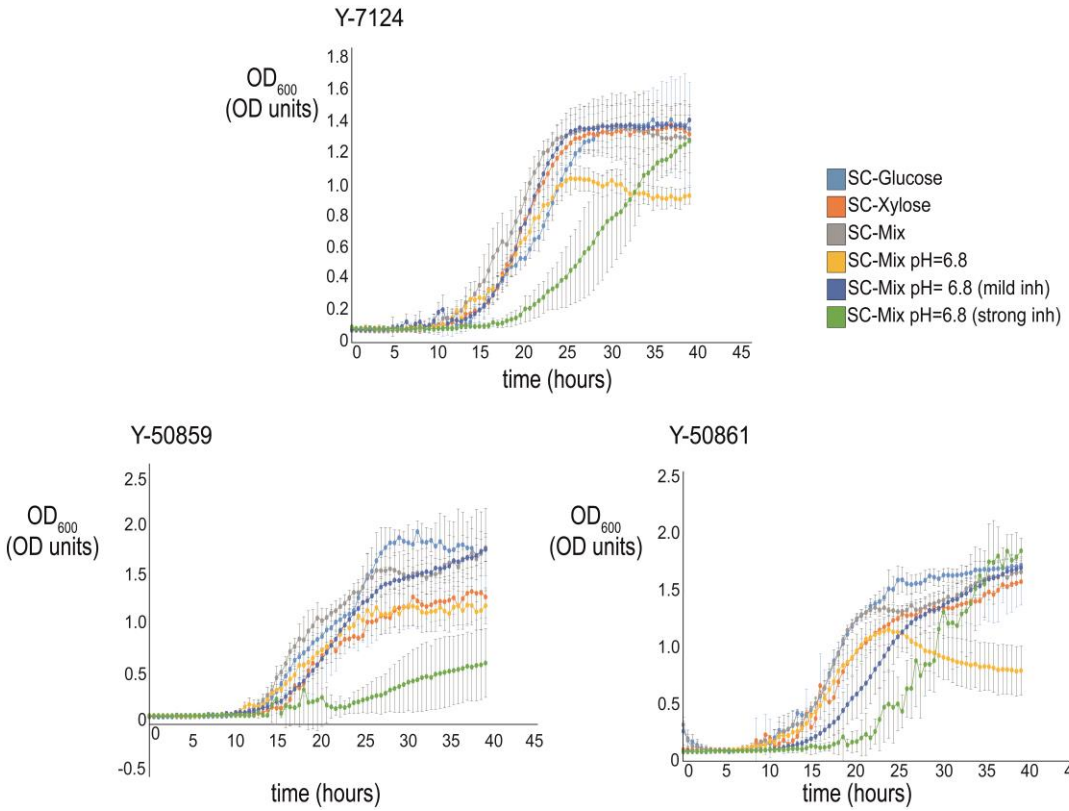


Figure 5.12. Growth curves of *S. stipitis* parental (Y-7124) and evolved strains (Y-50859 and Y-50861) in different carbon sources and in the presence of two different concentrations of a inhibitors cocktail (mild and strong inh). Growth is measured as change in OD_{600} (OD units) over time (hours). Experiments were conducted in 3 biological and 3 technical replicates.

No improvements were observed in the growth rate of the evolved strains when grown in minimal media using glucose or xylose as carbon source. However, the evolved strain Y-50859 showed higher growth rate compared to the parental Y-7124 when the carbon source was a mixture of glucose and xylose (**Figure 5.13, A**). This strain also exhibited higher maximum OD_{600} compared to Y-7124 in both SC-Glucose and SC-Mix medias (**Figure 5.13, B**), although not in SC-Xylose. Conversely, the evolved strain Y-50861 showed lower productivity compared to the parental strain in both SC-Glucose and SC-Mix, but not for SC-Xylose. Finally, both evolved strains show a delay in the lag time compared to the parental when the carbon source used was glucose. This effect is also observed respect to the strain Y-50859 in xylose, although this strain has a reduction in the lag time when the grown in SC-Mix (**Figure 5.13, C**).

Subsequently, the ability of the strains to grown in the presence of two different concentrations of an inhibitor mix in the SC-Mix media was assessed. This cocktail was prepared with some of the most common inhibitors found in lignocellulose hydrolysates, at its minimum (mild) and maximum (strong) reported concentration. To determine the inhibition effects of the mixture, the ratio of inhibitory and control conditions was calculated for the growth rate, maximum OD_{600} and lag time (**Figure 5.14**).

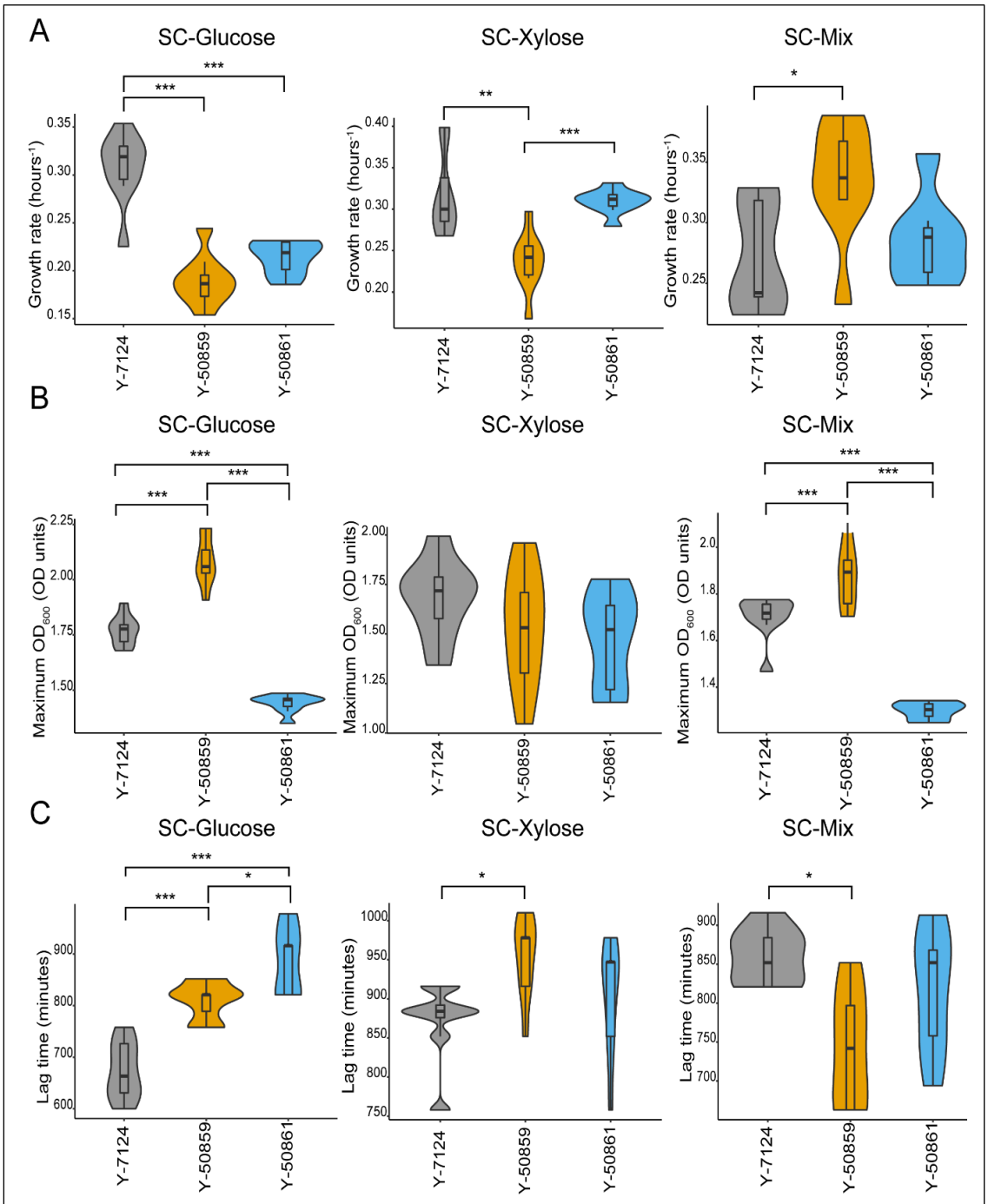


Figure 5.13. Growth rate (hours^{-1}) (A), maximum OD_{600} (OD units) (B) and lag time (minutes) (C) exhibited by the parental strain Y-7124 and the two evolved Y-50859 and Y-50861 in SC-Glucose, SC-xylose and SC-Mix. (*) indicate the level of significance.

Mild inhibitory conditions only decreased the growth rate of the parental strain, whereas for both evolved it was not affected. Surprisingly, the growth rate of the parental strain in strong inhibitory conditions was statistically the same than with no inhibitors, whereas it was lower for both evolved strains.

Both concentrations of inhibitory conditions affected the maximum OD_{600} in the three strains. The strains Y-7124 and Y-50861 showed an increase in the productivity in the presence of inhibitors, whereas the strain Y-50859 only increased it when the inhibitory conditions were mild, and it decreased under strong inhibition.

Finally, the lag time of the three strains is increased in strong inhibitory conditions, but in mild conditions it is only increased in Y-7124 and Y-50859, whereas it is not affected in Y-50861.

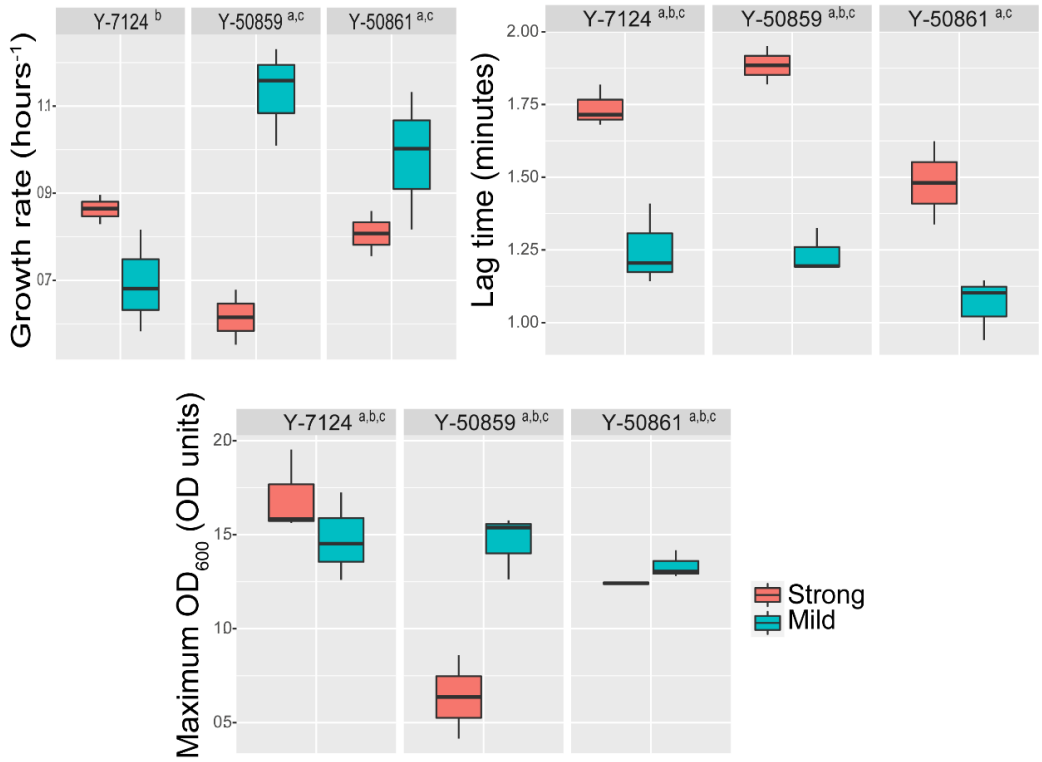


Figure 5.14. Ratios of the growth rate, maximum OD_{600} and lag time observed in mild (green) and strong (red) inhibitory conditions respect to their controls in each strain. Abbreviations: a: Statistical difference between strong inhibitory conditions and control. b: Statistical difference between mild inhibitory conditions and control. c: Statistical difference between mild and strong inhibitory conditions.

Apart from growth parameters under different carbon sources, two different biofilm related phenotypes were studied: agar invasion and cell sedimentation. To avoid variations caused by the differences in the growth observed previously, the sedimentation profiles were conducted starting with the same cell concentration (OD_{600}=1.0), and the agar invasion was normalised calculating the ratio of the cell invasion of the agar versus the cell growth on the surface of the plate, measuring therefore a relative invasion. The strains *S. cerevisiae BY4741* and *C. albicans SC5314* were used as negative and positive control of agar invasion. All experiments were conducted on SC-Mix media.

All the strains showed agar invasion, although none of them reached the levels of *C. albicans*. The strain Y-50861 showed more invasion than then parental strain Y-7124, whereas no differences were observed between the parental and Y-50859 (**Figure 5.15**). Finally, the strain Y-50859 showed improved sedimentation respect to its parental Y-7124, whereas no difference was observed respect to the strain Y-50861 (**Figure 5.16**).

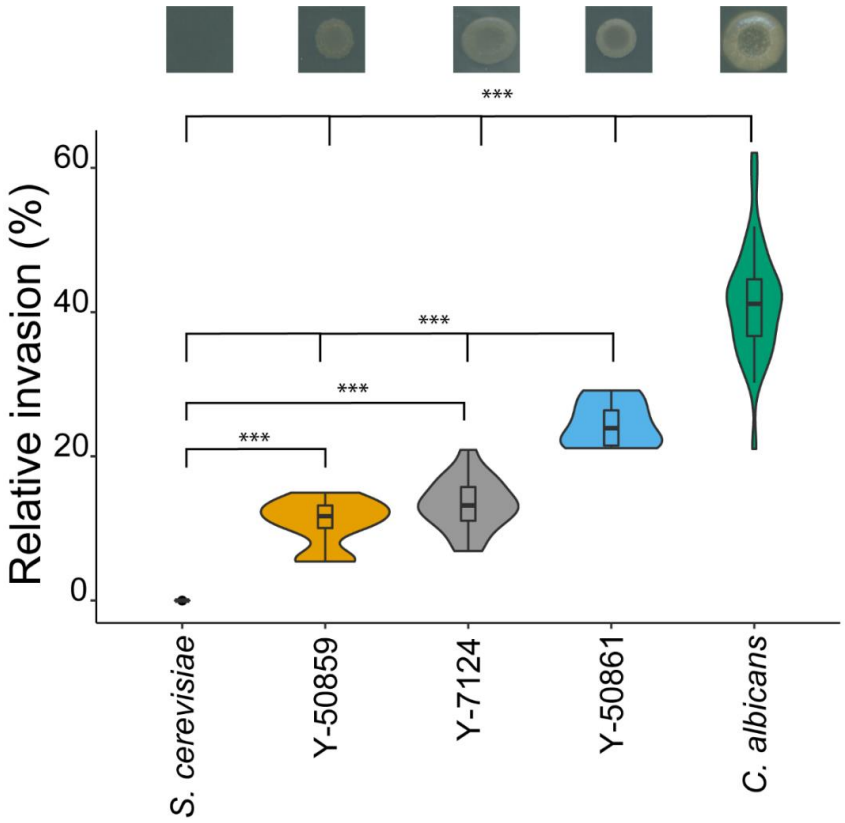


Figure 5.15. Percentage of relative agar invasion for the parental strain Y-7124 and evolved strains Y-50859 and Y-5061 in SC-Mix. The strains *S. cerevisiae BY4741* and *C. albicans SC5314* were introduced as negative and positive control, respectively.

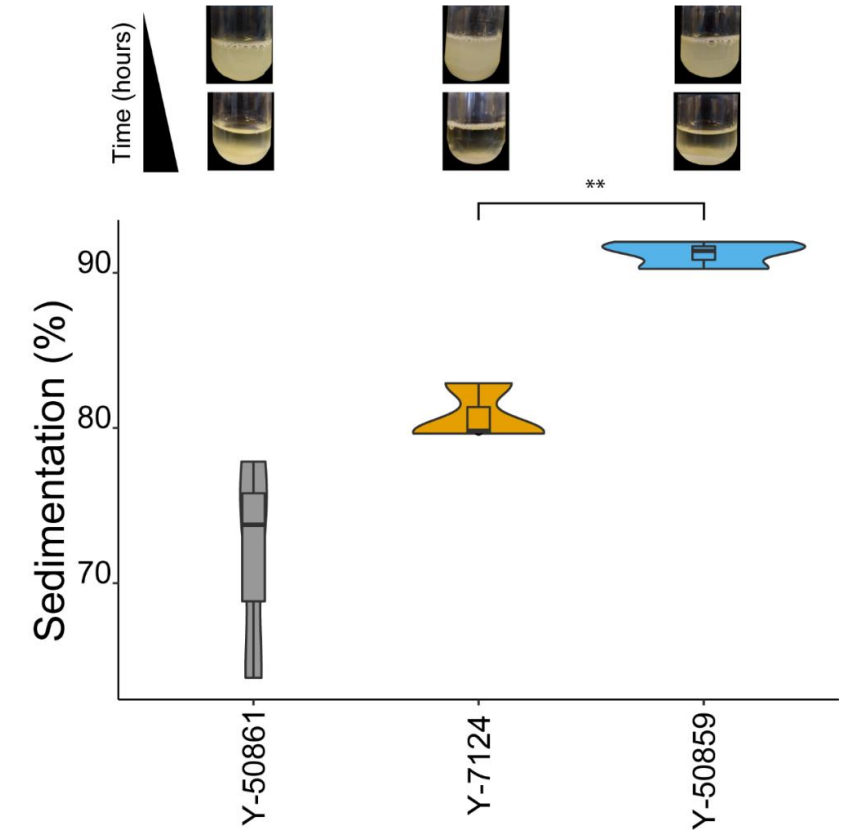


Figure 5.16. Percentage of sedimentation observed in the parental (Y-7124) and evolved strains (Y-50859 and Y-50861) after 4 hours in SC-Mix.

In summary, the adaptation of the strain Y-7124 to industrially relevant medias has caused modifications in the phenotypes of the evolved isolates (**Table 5.7**). The strain Y-50859, isolated from the adaptation of the parental Y-7124 to PSGHL exhibits an increase in the growth rate in SC-Mix and decrease in SC-Glucose and SC-Xylose, higher productivity in SC-Mix and SC-Glucose, higher lag time in SC-glucose and SC-xylose, and lower in SC-Mix, higher resistance to mild inhibition and lower to strong inhibition, and finally, higher sedimentation. Therefore, the only overall improvements observed in this study are in SC-Mix. On the other hand, the strain Y-50861, isolated from the adaptation of the strain Y-7124 to AFEX CSH shows lower growth rate in SC-Glucose and SC-Xylose, lower productivity in SC-Glucose and SC-Mix, higher lag time in SC-Glucose, similar resistance to inhibitors, and higher agar invasion, which means that no overall improvements are observed in any particular media among the studied.

Table 5.7. Effects of evolution on growth parameters studied in the evolved strains Y-50859 and Y-50861 respect their parental strain Y-7124. Abbreviations: +: Statistical improvement, - : Statistical detriment, N.E: No statistical effect.

Media	Parameter	Y-50859	Y-50861
SC-Glucose	Growth rate	-	-
	Maximum OD_{600}	+	+
	Lag time	-	-
SC-Xylose	Growth rate	-	-
	Maximum OD_{600}	N.E	N.E
	Lag time	-	N.E
SC-Mix	Growth rate	+	N.E
	Maximum OD_{600}	+	-
	Lag time	+	N.E.
	Sedimentation	+	N.E
	Agar invasion	N.E.	+
SC-Mix Inhibitors (mild)	Growth rate	+	+
	Maximum OD_{600}	N.E	N.E
	Lag time	N.E	N.E
SC-Mix Inhibitors (strong)	Growth rate	-	N.E.
	Maximum OD_{600}	-	-
	Lag time	N.E.	-

5.3 DISCUSION

Industrial fermentation conditions are stressful for microorganisms. Therefore, the use of yeasts that are able to grow and produce metabolites of interest in these conditions is important.

The production of ethanol using *S. stipitis* has been extensively studied in the last decades since it is the yeast with highest fermentation rate from xylose (Toivola *et al.* 1984). Although its performance compared to *S. cerevisiae* is still poor, several strains with improved fermentation phenotypes have been isolated, but the genetic drivers behind these changes remain unclear.

Hence, the goal of this section was the identification of the genetic changes in *S. stipitis* strains isolated after an adaptation experiment to two different industrially relevant medias, PSGHL (Y-50859) and AFEX CSH (Y-50861), and that exhibited improved fermentation phenotypes (Slininger *et al.* 2015).

According to their phenotypic characterization, the strain Y-50859 exhibited similar glucose uptake rate compared to the parental strain Y-7124, although the xylose uptake rate was higher, indicating a reduced diauxic trait. Moreover, when grown in optimal media (ODM), the evolved strain exhibited faster xylose fermentation than the parental in the presence of acetic acid when pre-grown in glucose, suggesting better resistance, although it was slower when pre-grown in xylose. When grown in SGH supplemented with nitrogen (SGH-N2), the evolved strain exhibits higher glucose and xylose uptake, higher ethanol productivity from both sugars, lower xylitol production and lower biomass productivity. Finally, no faster detoxification of furfural was observed compared to the parental, so the adaptation to this inhibitor might be through different mechanisms (Slininger *et al.* 2015). The results reported in this project indicate that this strain exhibits also exhibits a better sedimentation profile, and improved growth kinetics in SC-Mix (which might be related with the shortened diauxic time). On the other hand, the strain Y-50861 demonstrated higher xylose uptake than the parental when grown in AFEX CSH and in switchgrass hydrolysates supplemented with nitrogen (SGH-N1 and SGH-N2), but only showed improved ethanol production in SGH-N2. Moreover, the strain was more sensitive to acetic acid when grown in ODM (Slininger *et al.* 2015). Our results only indicated improvements in the growth under the presence of the inhibitors cocktail and in agar invasion.

Several modifications in the genome of the evolved strains respect to their parental have been detected in this study for both strains. These can be separated in two main groups: (i) modifications associated to repetitive DNA, and (ii) modifications non-associated to repetitive DNA.

The most important difference associated with repeats is the detection of a new chromosome (i-5L) in the evolved strain Y-50859, formed by repetitions of a fragment of the chromosome 5, and stable due to repetitive regions: the centromeres and the telomeres. This strain was isolated from the adaptation of the strain Y-7124 to PSGHL. This media is the liquor associated with switchgrass hydrolysates, which has been separated of the solids by filtration or centrifugation, and is characterised by high concentration of xylose, furan aldehydes and acetic acid, but low concentration of glucose and nitrogen. Aneuploidies caused by toxic stress has been described extensively in other yeasts. For example, the chromosome III in *S. cerevisiae* has been found to be completely and partially aneuploid in populations adapted to high levels of ethanol (Voordeckers *et al.* 2015; Morard *et al.* 2019). These same effects have been described in industrial *Saccharomyces* strains (Gorter de Vries, Pronk and Daran 2017), and also aneuploidies have been linked with survival in high salt and sugar environments in *Zygosaccharomyces rouxii* (Solieri, Dakal and Bicciato 2014).

Depending on the recombination class, cells use two main mechanisms for DSB repair: homologous recombination and non-homologous end joining (NHEJ). In homologous recombination, the 5'-terminal strand of a DSB is resected, freeing the 3' strand to invade and copy an intact homologous duplex. On the other hand, in NHEJ both ends of a DSB are captured and ligated with no homology required (Daley *et al.* 2005).

No homology was detected between the breakpoints in chromosome 1 and 2 of the strain Y-7124 that could explain a recombination dependant translocation, which suggests that the repair mechanism is non-homology dependent and occurred after a simultaneous break in two different genomic locations. This is supported by previous studies that suggest that NHEJ is the predominant repair pathway in *S. stipitis* (Maassen *et al.* 2008). Despite of this, Maassen *et al*. also reported that homologous recombination is still active and efficient.

DSBs created at centromere are normally repaired either by gene conversion or by BIR. The repair by BIR between the centromere repeats from a chromatid can lead to the formation of an isochromosome. We hypothesise a DSBs happened at the centromere of the strain Y-50859 during the S phase and this was repaired by BIR (**Figure 5.17**). This theory is supported by previous observations of template switching during BIR (Smith, Llorente and Symington 2007). Isochromosomes formed by recombination between inverted repeats in the centromere region have been observed previously in *S. pombe* and in *C. albicans* (Selmecki, Forche and Berman 2006; Nakamura *et al.* 2008; Tinline-Purvis *et al.* 2009). Consistently, analysis of the centromeric sequence of chromosome 5 in Y-50859 identified inverted repeats as its

main structure (several Tps5a pseudotransposons are found in the central region of the centromere, and two Tps5b are found in the exterior [**Figure 5.11, B**]).

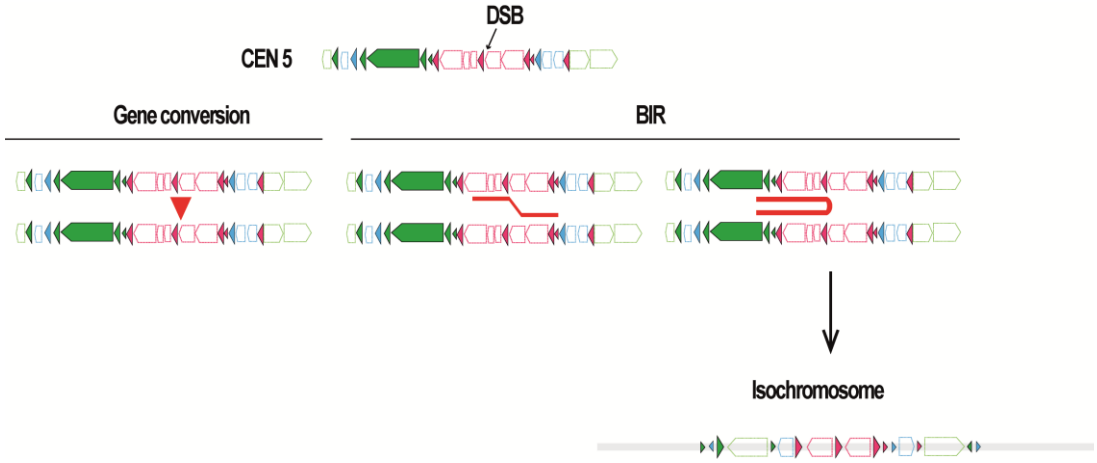


Figure 5.17. Model for DSB repair at centromeric repeats and their relationship with isochromosome formation.

Onaka *et al.* demonstrated that Rad51 and Rad54 promoted non-crossover recombination between centromere repeats on the same chromatid to prevent the formation of isochromosomes in *S. pombe* by observing an increase in the spontaneous isochromosome formation in *rad51Δ-rad54Δ* mutants (Onaka *et al.* 2016). Both Rad51 and Rad54 were identified in *S. stipitis* during this study by homology with *C. albicans* proteins, although no point mutations were detected in the strain Y-50859 respect to either Y-7124 or Y-11545, which is consistent with the low frequency in the detection of isochromosomes in the strains studied in this project.

Moreover, the stability of the new chromosomes that derive in aneuploidies depends also in the presence of telomeres. Telomeric sequences were detected in the nanopore sequencing reads belonging to the isochromosome, and its presence was further confirmed by PCR (**Figure 5.11, D**). As previously demonstrated in *S. cerevisiae,* when a segment containing an active centromere is duplicated, DNA ends can acquire telomeres by initiating a recombination-dependent DNA replication (McEachern and Haber 2006). A similar system has also been identified in *C. galbrata* (Poláková *et al.* 2009).

A total of 26 genes have been predicted in this chromosome (**Table 5.6**), but none of them is essential for survival in rich medias since we have detected a loss of i-5L in 4% of the cells according to fluctuation analysis. Structure analyses indicate that the aneuploidy is an isochromosome (a chromosome in which the two arms are a mirror of each other), which means that these genes are triplicated (one copy in the chromosome 5 and two in the isochromosome).

Therefore, overexpression of these genes could be expected, and this can be related with the different phenotypes previously described in the strain Y-50859.

First, it is noticeable that several genes in the region are involved in DNA repair and stress responses (*PICST_89662 (PRE4)* (Wang *et al.* 2008; Galanty *et al.* 2012)) and *PICST_46961 (RPN8* (Galanty *et al.* 2012)), and in the maintenance of the kinetochore structure (*PICST_89595 (CAC2)* (Sharp *et al.* 2002)), which might be advantageous considering the large chromosomal modifications detected in this strain (aneuploidy and reciprocal translocation).

Multiple genes are related to sugar metabolism and therefore possibly related with improvements in fermentation. Higher sugar uptake can be related with overexpression of β-N-acetylhexosaminidase (*PICST_32069*), which acts as carbon scavenger in *C. albicans* (Ruhela *et al.* 2015), but also in the degradation of oligosaccharides and in the remodelling of the cell walls in filamentous fungi, such as *Aspergillu oryzae* (Plíhal *et al.* 2007). Remodelling in the cell wall was observed in growing cells of the strain Y-50859, since their shape was more filamentous than in the parental strain (data not shown). This can be also related with the overexpression of the gene *PICST_46807* (*DOP1*), a Dopey leucine zipper transcription factor, related to hyphal morphogenesis in *A. nidulans* (*DOPA*) and in *S. cerevisiae* (Pascon and Miller 2000). The hyphal growth has also been related with cell-to-cell adhesion (Alsteens *et al.* 2013), which might then condition the better sedimentation profile we have observed in this evolved strain (Stewart 2018).

The strain Y-50859 exhibited less biomass productivity in both phenotype experiments under consideration (in SGH-N2 conducted by Slininger *et al.* (Slininger *et al.* 2015) and in SC-glucose and SC-xylose in this study). The overexpression of lysophospholipases (*PICST_83761*) in *S. cerevisiae* is related with the accumulation of palmitic acid (De Smet *et al.* 2013), which is at the same time related with growth decrease (Nozawa *et al.* 2002). Despite of this, the lower biomass productivity observed in this strain might be related to different causes. First, growth defects have been extensively described in aneuploids strains of *S. cerevisiae*, mainly to a delay at the G1 phase in mitosis caused by larger DNA content (Torres *et al.* 2007; Pavelka *et al.* 2010). Moreover, lower biomass productivity is commonly reported among strains isolated after ALE experiments (Jasmin, Dillon and Zeyl 2012; van Dijk *et al.* 2019), probably due to shifting substrates processing from growth to metabolite production.

Finally, heme A farnesyltransferases (*PICST_47531, COX10)* are required for the activity of the cytochrome c oxidase (cox1), the last subunit of the respiratory chain that catalyses the reduction from oxygen to water. *COX10* overexpression in *C. albicans* elevates the levels of newly synthesise cox1 and restores deficiencies in respiration (Pierrel *et al.* 2008). Cox1 expression has also been related to partial respiration rescue

in defective mutants of *S. cerevisiae* (Barrientos 2002). Fermentation in *S. stipitis* is carried out in oxygen limited conditions. Therefore, overexpression of *PICST_47531* might collaborate in scavenging oxygen and in a more efficient respiration system.

Interestingly, the gene *PICST_65754* codifies for a predicted threonine synthase. Overexpression of this enzyme in *S. cerevisiae* is related with an increase in the cellular concentration of 2-ketobutyrate (Nishimura *et al.* 2018), an intermediate for 1-propanol production and also related to the production of higher chain alcohols with potential use in the biofuel industry (Atsumi, Hanai and Liao 2008). To our knowledge, no studies have been conducted on propanol and higher chain alcohols production in *S. stipitis.*

Apart from the aneuploidy, the modifications associated to repetitive regions are mainly related to centromeres. Although the average size of the centromeres is maintained in the three strains, the organization in 3 of them has changed. First, the centromere of chromosome 1 in the evolved strain Y-50861 is approximately 4.0 kb larger due to the presence of an extra LARD. Moreover, the centromere of chromosome 5 of both evolved strains is 5.0 kb larger due to the presence of an extra Tps5a. Finally, the centromere of chromosome 7 of the parental strain Y-7124 is 7.0 kb larger than in the two evolved due to the presence of an extra Tps5b.

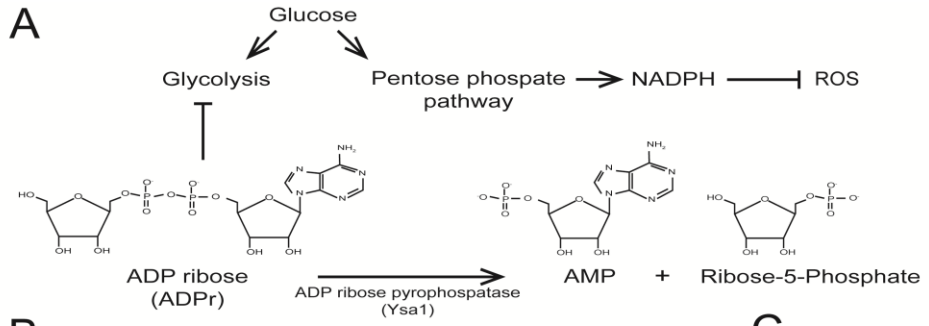
Considering that specific centromeric DNA sequences are not essential for the function of the centromere, it is not surprising that centromeres are rapidly evolving in eukaryotic genomes (Csink and Henikoff 1998; Murphy and Karpen 1998; Rosin and Mellone 2017). Phylogenetic analysis of candidate centromeric repeats from 282 animal and plant species have revealed little sequence homology over 50 million years of divergence (Melters *et al.* 2013), although there are examples of conserved centromeric sequences in mouse and humans (Earnshaw *et al.* 1987). These variations are normally associated with unequal crossing over during meiosis I, strand slippage during DNA replication, or, like probably in this case, the transposition of mobile DNA elements (Rosin and Mellone 2017). Polymorphisms in the centromeres of different isolates have been previously reported in yeasts such as *Komagataella phaffi* (formerly called *Pichia pastoris*) (Coughlan *et al.* 2016) or *Candida parapsilosis* (Ola *et al.* 2020).

On the other hand, other repetitive regions that showed variations between the two natural isolates did not show modifications after adaptation, since the gene families and distribution of genes in the subtelomeric regions and the number and distribution of non-centromeric transposable elements were maintained in the two evolved strains respect to the parental.

The principal difference between the parental and evolved strains unrelated to repetitive DNA is the reciprocal translocation between chromosome 1 and 2 in the strain Y-50859. Similar rearrangements caused by exposition to environmental stresses in industrial strains have been previously described (Codón, Benítez and Korhola 1997;

Rachidi, Barre and Blondin 1999; Querol *et al.* 2003; Chang *et al.* 2013). The translocation observed in Y-50859 is associated with the disruption of an unannotated ORF, codified by the gene *PICST_53805*.

Homology studies of the ORF sequence allowed their identification as a putative ADP-ribose pyrophosphatase (*YSA1*). This enzyme is the responsible of the transformation of ADP-ribose (ADPr) to AMP and D-ribose-5-phosphate (**Figure 5.18, B**), and has been identified in yeasts such as *S. cerevisiae* (Dunn *et al.* 1999) or *Cryptococcus neoformans* (Lee *et al.* 2014).



B

S. stipitis
C. abicans
S. cerevisiae
A. nidulans

C

Hydrolase	Y-7124	Y-50859
PICST_88634	D	D
PICST_53805 (Ysa1)	D	N.D.
PICST_36667	D	D
PICST_79554	D	D
Npy1	D	D
Ddp1	D	D
Tnr3	D	D
Idi1	D	D

D

Figure 5.18. Ysa1 enzymatic reaction and model on how ROS production is inhibited (Adapted from Tong *et al*. (2009) (A). Alignment of different fungal Ysa1 proteins (B) Table with all the proteins identified in *S. stipitis* with a NUDIX domain (C) Alignment of the different proteins with a NUDIX domain in *S. stipitis* (D). Abbreviations: D: Detected, N.D: Not detected.

Ysa1 presents a nucleoside diphosphate linked to moiety-X (NUDIX) hydrolases domain, or Nudix box, a highly conserved 23 residue motif (GX5EX7REUXEEXGU, where U is a bulky hydrophobic residue and X is any residue) (Mildvan *et al.* 2005) (**Figure 5.18, C**). These hydrolases are characteristic for their high substrate variability, and hence are involved in the hydrolysis of different types of compounds, mainly canonical (d)NTPs, oxidised (d)NTPs, non-nucleoside polyphosphates and capped mRNAs (Carreras-Puigvert *et al.* 2017). Nevertheless, the biological functions of many of these hydrolases is yet unclear.

Interestingly, phenotype studies conducted in *S. cerevisiae* correlate the disruption of *Ysa1* with an increase in the resistance to oxidative stress. This is related to the accumulation of the ADPr not processed in *S. cerevisiae Δysa1* cells. High levels of ADPr inhibit glyceraldehyde-3-phosphate dehydrogenase, which switches the metabolic processing of glucose from glycolysis to the pentose phosphate pathway. This ultimately leads to the accumulation of NADPH, responsible of the increased resistance to oxidative stresses (Tong, Lee and Denu 2009) (**Figure 5.18, A**). Similar protective effects have also been observed in *C. neoformans* (Lee *et al.* 2014).

Nevertheless, enzymes from the NUDIX superfamily present substrate variability, and therefore alternative member of the family might act to compensate the activity of the disrupted Ysa1. Subsequently, all the NUDIX enzymes present in *S. stipitis* genome were identified and their sequence was compared by Clustal-Wallis alignment.

A total of 8 NUDIX hydrolases were identified in the *S. stipitis* genome (**Figure 5.18, C).** All of them, except Ysa1 were present in both strains Y-7124 and Y-50859. However, none of the enzymes identified presents sequence similarity to the truncated Ysa1 (**Figure 5.18, D**). Therefore, we conclude that the activity of ADP-ribose pyrophosphatase is likely to be lost.

The strain Y-50859 was isolated after the adaptation of the *S. stipitis* natural isolate Y-7124 to PSGHL, a media with reported: (i) high concentration of xylose (pentose) and low concentration of glucose (ii) high concentration of inhibitors released after lignocellulose processing (Slininger *et al.* 2015). We hypothesise that the truncation of a gene that causes the switch of metabolism towards pentoses, and provides higher resistance to oxidative stress seems advantageous, and therefore likely to have been selected during the adaptation.

Apart from the translocation, this study identified 3 missense SNPs in the strain Y-50859, whereas 11 missense and 1 nonsense were detected in the strain Y-50861 (**Table 5.2**). However, none of the modifications observed in the phenotypes seems related to these SNPs. The reason why more SNPs are present in the strain Y-50861 might be related with the use of two different medias during the ALE experiment (AFEX CSH and ODM with high concentration of ethanol) (Slininger *et al.* 2015).

Curiously, the 3 missense SNPs observed in Y-50859 were also observed in Y-50861. The genes affected were *PICST_46361* (*OCA1* in *S. cerevisiae*, a protein tyrosine/serine phosphatase), required for the response against oxidative stress (Alic, Higgins and Dawes 2001), *PICST_59269*, a transporter of the major facilitator superfamily, and *PICST_31552* (a *PAP-1* (poly(A) polymerase), essential for polyadenylation (Minvielle-Sebastia *et al.* 1997)). No domain modifications have been detected in any of the proteins by SMART analysis, and therefore they might still be functional.

The strain Y-50861 presents 10 additional SNPs (1 synonymous, 8 missense and 1 nonsense). The gene *PICST_56002*, which codifies for a GTP-binding ADP ribosylation factor (*ARF1* in *S. cerevisiae*) is associated with polyadenylate-binding proteins recruitment at the Golgi complex (Stearns *et al.* 1990; Trautwein *et al.* 2004), and presents the accumulation of 4 missense SNPs and the only nonsense mutation reported in the evolved strains. Moreover, the genes *PICST_62836* and *PICST_73970*, which codify for an MCP-domain (membrane compartment of Pma1, which is an essential ATPase in the plasma membrane of the cell (Serrano, Kielland-Brandt and Fink 1986)) signal transduction protein and for a predicted P-loop ATPase fused to an acetyltransferase, are also affected by missense SNPs (**Table 5.3**). The rest of SNPs affect to genes whose protein product has not been characterised. No modifications in the domains of these proteins were observed either by SMART analysis.

5.4 CONCLUSIONS

This section aimed to study whether the genome of strains isolated from the parental *S. stipitis* strain NRRL Y-7124 through an ALE experiment, and with improved phenotypes, reported in this and in previous studies, had suffered any modifications associated to genome instability.

The karyotype of 4 evolved strains was studied by CHEF electrophoresis and compared with their parental. Only one of the evolved strains (Y-50859) showed modifications in the banding profile. Therefore, the genome of the strains Y-50859 and Y-50861 (as a control with no genome modifications), were sequenced using a hybrid TGS approach: a combination of Oxford Nanopore technologies (ONT) and Illumina.

Several modifications were detected and classified depending on their association or not to repetitive DNA regions.

The main modification observed associated to repetitions was the detection of an aneuploidy in the evolved strain Y-50859. This aneuploidy consists in an extra isochromosome (i-5L) formed by repetitions of a region of the chromosome 5, and that is also stable thanks to repetitive regions: the centromere and telomeres. Several of the

genes present in this region are candidates to be responsible of the improvements described in the phenotype of the strain Y-50859. Moreover, modifications in the centromere structure were detected for both strains. This result is a confirmation that CHEF electrophoresis does not provide enough resolution when modifications are small, since no alterations were detected in the karyotype of the strain Y-50861. No variations were detected in the number and distribution of transposable elements or genes present in the subtelomeres, despite differences reported in this study between two different natural isolates (Y-11545 and Y7124, see chapter 3).

A reciprocal translocation not associated with repetitive DNA was also detected between chromosomes 1 and 2 of the evolved strain Y-50859. This translocation caused the disruption of the gene *YSA1*, whose activity is likely to be lost. We hypothesise that the loss of this activity switches the metabolism of the strain from glycolysis to the pentose phosphate pathway, which allows simultaneously: (i) Improvements on the growth in the xylose-rich media in which the strain was adapted (ii) Accumulation of NADPH, which enhanced resistance to environmental stress.

Moreover, several SNPs were detected between the parental strain Y-7124 and the evolved strains Y-50849 (3) and Y-50861 (12), although none of them seems related with the improvements in the phenotypes reported for this strain in the current or in previous studies.

5.5 FUTURE WORK

This study has detected an aneuploidy in the strain *S. stipitis* Y-50859, isolated from the adaptation of the strain Y-7124 to PGSHL. This aneuploidy is due to an extra isochromosome (i-5L) formed of repetitions of a region of the chromosome 5. There are several genes in the chromosome that might explain the improved phenotypes observed in the strain. During this study, strains of Y-50859 with loss of i-5L have been isolated. Therefore, the new isolated strains can be studied to link improvements with the presence of the i-5L.

Moreover, this strain also suffered the disruption of the gene *YSA1* by a reciprocal translocation. However, comparisons between the phenotypes observed in *S. cerevisiae* Δ*ysa1* cells and Y-50859 were not conducted.

Moreover, the reintroduction of *YSA1* in the strain Y-50859 or its disruption in Y-7124 through genetic engineering techniques could elucidate whether modifications in the phenotypes are more associated to the disruption of the gene or to the presence of the i-5L.

No gene expression studies were conducted in this project, and therefore, although several genes have been proposed as responsible of the improved phenotypes, their actual effects on the cell, or even functionality has not been studied yet.

Finally, this study has confirmed that not all genome modifications are translated in karyotype alterations. Therefore, the study of the genome of the other two evolved strains that did not show variations in the chromosome banding profile might be interesting to shed some light in the specific mutations that cause phenotype improvements.

Chapter 6

Discussion

In this project, we have analysed for the first time the genome plasticity of *S. stitipis*, a yeast that belongs to the CTG clade, and that is one of the most promising organisms for bioethanol production.

Our results clearly demonstrate that the genome conformation of natural isolates of *S. stipitis* is widely variable. Genome plasticity has been widely studied in other yeasts belonging to the CTG clade, such as *C. albicans,* a common human pathogen, where it has been associated with its rapid adaptation to physiological changes in the host and to stresses derived from antibiotic treatments (Selmecki, Forche and Berman 2010; Legrand *et al.* 2019; Todd *et al.* 2019). Nevertheless, plastic genomes have also been reported in yeasts that do not belong to the CTG clade. For example, extensive karyotype variations have been described in natural isolates of *S. cerevisiae* (Sniegowski, Dombrowski and Fingerman 2002).

6.1 The genome of *S. stipitis*: presence of repetitive regions

Repetitive regions have been previously described as hotspots of genome plasticity since they might lead to instability and genome shuffling under stressful conditions (Freire-Benéitez *et al.* 2016; Todd *et al.* 2019; Dunn and Anderson 2019). There are two main reasons behind this: (i) Repetitive regions are prone to recombination, (Argueso *et al.* 2008; Hoang *et al.* 2010) (ii) Non-canonical DNA structures are commonly observed in repetitive loci, which can lead to a collapse of the replication fork and to double strand breaks (Aguilera and García-Muse 2012).

In this study we have identified three main repetitive elements in the genome of *S. stipitis*: transposable elements, the chromosome ends (telomeres and subtelomeres), and the centromeres, which had previously been characterised by several authors (Lynch *et al.* 2010; Cao, Gao, *et al.* 2017; Cao, Seetharam, *et al.* 2017; Coughlan and Wolfe 2019).

Transposable elements are the main contributors to repetitive DNA in the genome of *S. stipitis*. We have identified 6 families of Class I transposable elements, 3 of them LTR-transposons (*Ava, Bea* and *Caia*), and 3 non-LTR (*Ace, Bri* and *Can*), of which two, *Ace* and *Can* have been described in yeasts for the first time. All 6 families exhibit similar basic structure commonly described in transposable elements: 2 ORFs and surrounding non-coding regions (Bourque *et al.* 2018).

Our results demonstrate that *Ace* and *Can* are phylogenetically close, and that they belong to the L1 family of LINE non-LTR transposable elements. Despite the high

sequence similarity between their POL, the second ORF present in the *Ace* and *Can* (LINEA1 and LINEA3, respectively) exhibit higher variability. Homology studies by BLASTP report a 30% of amino acid sequence similarity with the POL detected in *Metschnikowia aff pulcherrima*, which suggest that it is quite different from the closest homologue in other yeasts species. On the other hand, *Bri* has been previously described as the Zorro3 transposon in *C. albicans* (Goodwin, Ormandy and Poulter 2001), which suggests that their presence in the genome is prior to the species diversion from a common ancestor.

The three families described as LTR transposons (*Ava, Bea* and *Caia*) exhibit the same structure as other LTR elements described in yeasts. However, no Gag protein was detected in this study. Instead, a specific protein to *S. stipitis* was discovered associated to the elements, which we named LAP (LTR-associated protein). Since strong variability between Gag proteins have been reported for *S. cerevisiae* (Jordan and McDonald 1998; Kim *et al.* 1998; Bleykasten-Grosshans, Friedrich and Schacherer 2013), we hypothesise that these proteins act as Gag in *S. stipitis.*

This study has identified the telomeric repeats in *S. stipitis* for the first time. These are different from the canonical telomeres, since they are formed by tandem repeats of 24bp, instead of the canonical 6bp GT motif. Despite of this, there is a T+G motif conserved respect to typical telomeric repeats.

Finally, the subtelomeric regions (<30 kb distance from telomeric repeats) have also been described for the first time in this study. Subtelomeres are characterised by low gene density and epigenetic silencing (Pryde and Louis 1999). Moreover, these regions have been reported as repetitive in the genome of several organisms and therefore difficult to assemble (Sadhu *et al.* 1991; Louis 1995; Vershinin, Schwarzacher and Heslop-Harrison 1995; Torres *et al.* 2011). Despite of this, we have detected multiple copies of genes of different families, commonly described as subtelomeric genes: DEAD helicases (from the aminoacids forming the characteristic motif in this enzyme (Asp, Glu, Ala, Asp or DEAD)), major facilitator superfamily of transporters and fungal transcription factors. The presence of specific gene families that might condition phenotypes has been reported in the subtelomeric regions of different organisms. For example, genes involved in carbohydrate metabolism and in biofilm formation have been reported in yeasts subtelomeres (Michels and Needleman 1984; Carlson, Celenza and Eng 1985; Naumov *et al.* 1990).

6.2 The genome plasticity of *S. stipitis*: effects of repetitions

Our karyotype analysis of natural isolates of *S. stipitis* obtained from different habitats and countries demonstrates how remarkably variable its genome is.

However, the sequencing of a second natural isolate (Y-7124) was required to prove that the genome of *S. stipitis* is plastic. This allowed us to discover differences in several genomic regions between the two isolates.

The major difference detected was in the number and distribution of transposable elements. Although the same 6 families were present in both strains, the total number of full-length transposons in Y-11545 was 45 (which accounts for a 2% of the total genome content), whereas it was 20 for Y-7124 (1% of the total genome size). From these, only 10 loci (~15%) were present in both strains, which suggests that they are elements that rarely transpose. The rest of the transposons (~85%) are present in only one of the strains, which leads us to hypothesise that they might still be active and competent of transposition. However, further studies would be required. For example, the levels of expression and activity of the reverse transcriptase present in Zorro3 elements in *C. albicans* has been conducted by Northern blot (Deininger and Belancio 2016).

In addition, the two strains exhibit differences in total size of the centromeres and in the distribution of their associated repetitive elements. The strain Y-11545 presents larger centromeres, and they were richer in centromeric transposons and transposon-derived repetitions. This confirms previous observations that epigenetically defined centromeres in yeasts are highly affected by evolutionary forces (Padmanabhan *et al.* 2008).

Differences in the number and orientations of genes present in the subtelomeres were also detected between the two natural isolates of *S. stipitis*, which is a suggestion of increased rate of recombination and mutation. This is an indication that variability in subtelomeres caused by instability and gene recombination might condition adaptation to novel niches, which has been previously reported in several ascomycetes (Brown, Murray and Verstrepen 2010). Moreover, high plasticity in the subtelomeric regions has also been described several organisms, from bacteria to humans (Yue *et al.* 2017; Ricchetti, Dujon and Fairhead 2003; Thibessard and Leblond 2014; Flint *et al.* 1995).

Finally, the comparison between the two genomes allowed the identification of one reciprocal translocation between chromosome 5 and 7, which explains the differences observed in the karyotype. One of the breakpoints of this translocation was present in a 136 kb region rich in repetitions (both related and non-related to

transposons) in the chromosome 7 of the strain Y-11545. Additionally, a 20 kb fragment formed by inverted repeats was present in this repetition-rich region in the chromosome 7, but it is missing in the strain Y-7124. This has been reported for different organisms. For example, a 280 kb region that acts as a transposon hotspots has been described previously in maize (SanMiguel *et al.* 1996) and inverted repeats of transposable elements are recombination hotspots in *S. cerevisiae* (Charles and Petes 2013). Despite of the detection of a reciprocal translocation associated to repetitive elements between the two strains, their evolutionary history is unknown, and therefore, the association of repeats with inherent instability in the genome is complicated.

Nevertheless, we demonstrated that the genome of *S. stipitis* is unstable, since extensive karyotypes modifications resulted from an ALE experiment to temperature stress. Similar effects have been observed in other yeasts (Yona *et al.* 2012; Sirr *et al.* 2015). Surprisingly, control conditions in minimal media offered higher variability than temperature stress, which suggests that: (i) The genome of *S. stipitis* is inheritably instable even in non-stressful conditions, or (ii) Minimal media is more stressful for *S. stipitis* growth than anticipated. Since no karyotype variations were observed after routine growth in YPD we hypothesise that minimal media is potentially the responsible for this instability. Increases in the chromosome alterations with depletion of nutrients have also been reported in *C. albicans* (Rustchenko 2007).

6.3 Genome plasticity and phenotypic variations in *S. stipitis*

Despite the major differences detected in the genome organization of the two strains, the number and function of predicted genes is conserved, which suggests that little phenotypic variation should be expected. Our analysis confirms this hypothesis, since the overall growth kinetics studied for the two strains is similar, although the strain Y-7124 showed slower growth rate in SC-Mix, lower productivity in SC-Glucose and SC-Mix, shorter lag time in SC-Glucose, and higher sedimentation when compared to the strain Y-11545.

Despite this, several studies have demonstrated the link between differences in phenotypic behaviour and genome organizations (Hirakawa *et al.* 2015). This is also observed in this study, since natural isolates that exhibit the same karyotype present similar phenotypic behaviours. Remarkably, the strains that belong to the genomic organization groups G4, G5, G6 and G7 are the best performers, which is surprising, since the most used strains in *S. stipitis* research belong to the groups G1 (Y-11545) and G3 (Y-7124). Therefore, we propose that the use of other strains with better phenotypic traits could be interesting to further increase the potential use of *S.*

stipitis at industrial level. Nevertheless, more extensive research about fermentation yields and scale up adaptability in these strains is required.

6.4 Genome plasticity and repetitions in *in vitro* evolved strains of *S. stipitis*

Finally, we have demonstrated that genome plasticity is also observed in *S. stipitis* strains with improved fermentation phenotypes after *in vitro* adaptation to industrially relevant medias. The strain Y-50859 presented karyotype modifications with respect to its parental Y-7124, and these are both linked and non-linked to repetitive DNA. Extensive karyotype variations have also been described in *S. cerevisiae* strains that exhibited improved fermentation phenotypes after isolation from adaptation experiments (Nadal *et al.* 1999; Carro and Piña 2001; Sniegowski, Dombrowski and Fingerman 2002).

The most important difference associated to repetitions was the detection of aneuploidies associated with the presence of a ~180 Kb extra isochromosome (i-5L), formed by repetitions of a region of chromosome 5, and whose stability depends on repetitive sequences: the telomeres and centromere. Aneuploidies and CCNVs caused by toxic stress have been described extensively in other yeasts (Voordeckers *et al.* 2015; Morard *et al.* 2019; Gorter de Vries, Pronk and Daran 2017; Solieri, Dakal and Bicciato 2014).

These aneuploidies are normally caused by missegregation of the chromosomes during either mitosis or meiosis, supported by errors in the checkpoints of the cell cycle and division. Several factors, like chemical stress or ethanol concentration in the media have been proven to increase chromosome missegregation (Gorter de Vries, Pronk and Daran 2017). However, in this case the CCNV detected is not from a complete chromosome, but from a region of it. As previously stated, inverted repeats act as mitotic recombination hotspots in yeasts (Charles and Petes 2013). The centromere of chromosome 5 in Y-7124 (whose duplication originated the i-5L chromosome) is characterised by large inverted repeats related to centromeric transposons. Therefore, we hypothesise that these repeats might have acted as a mitotic recombination hotspot.

The presence of aneuploidies has normally detriment effects on the cells. First, they are a source of genome instability, so cells that carry an aneuploidy present more probability of accumulating other genome modifications (Jason M. Sheltzer *et al.* 2011). Second, the overexpression of genes can lead to the accumulation of misfolded proteins, which might be responsible of proteotoxic stress (Oromendia, Dodgson and Amon 2012). Finally, growth can be affected by several reasons: (i) Aneuploidies are normally

associated with the upregulation of genes involved in stress responses and downregulations of genes involved in growth (Torres *et al.* 2007) (ii) Higher energy is consumed for the overproduction of protein and their correct folding, so there is a higher proportion of nutrients that are not used for cell growth (Kerry A. Geiler-Samerotte *et al.* 2011). In accordance, defects in the growth were observed in the aneuploid strain when grown in SC-Glucose, SC-Xylose, and also in YPD (data not shown).

Considering the possible effects that aneuploid yeast exhibit, controlling the gene expression of the extra genes seems a good strategy to limit the possible reductions in fitness. Hose *et al.* (Hose *et al.* 2015) have demonstrated that cells that carry several copies of a gene, express those genes less often than expected, a phenomenon called dosage compensation, which might reduce the negative effects of the duplications while increasing the adaptability to new niches.

Therefore, we propose that (i) The genes carried in the i-5L isochromosome (25) are probably not expressed at optimal levels, which can be assessed by qPCR (ii) The expression of the genes in chromosome i-5L offer selective advantages in the stressful conditions in which the strain was adapted, and therefore, their maintenance will be conditioned to these conditions. This is suggested by the detection of a ~4% of chromosome loss by fluctuation assays in non-stressful conditions, which also indicates that the isochromosome is not required for survival. However, fluctuation analysis in PSGHL to compare the chromosome loss levels with the results observed in YPD is necessary to demonstrate our hypothesis.

Apart from changes in the ploidy level, a change in the distribution and size of the centromeres in two chromosomes was also detected (centromere 5 was larger in the evolved strain, whereas centromere 7 was larger in the parental strain). These effects were also observed in the other evolved strain (Y-50861). Such polymorphisms in the centromeres of different isolates have been previously reported in yeasts (Coughlan *et al.* 2016; Ola *et al.* 2020).

The main difference not associated with repetitive regions between the strains is a reciprocal translocation between chromosome 1 and 2. Similar rearrangements caused by exposition to environmental stresses in industrial strains have been previously described (Codón, Benítez and Korhola 1997; Rachidi, Barre and Blondin 1999; Querol *et al.* 2003; Chang *et al.* 2013). This translocation is associated with the disruption of the gene *YSA1,* responsible of the transformation of ADP-ribose (ADPr) to AMP and D-ribose-5-phosphate. The deletion of *YSA1* in *S. cerevisiae* increases the resistance to

oxidative stresses via the accumulation of NADPH through a switch in sugar metabolism from glycolysis to the pentose phosphate pathway (Tong, Lee and Denu 2009).

The strain Y-50859 was isolated after the adaptation of the *S. stipitis* natural isolate Y-7124 to PSGHL, a media with reported: (i) high concentration of xylose (pentose) and low concentration of glucose (ii) high concentration of inhibitors released after lignocellulose processing (Slininger *et al.* 2015). Therefore, we hypothesise that the truncation of this gene is advantageous, and therefore likely to have been selected during the adaptation.

Although this project did not study the direct correlation between the genomic changes and the improved fermentation phenotypes observed in the *in vitro* adapted strains, it is strongly suggested that genome alterations have conditioned phenotypic traits related to bioethanol production. Further phenotyping of the strains that have lost the isochromosome and that have been isolated during this study will shed a light in its effect on the phenotypic diversity of *S. stipitis.*

Chapter 7

Conclusions

During this study, karyotype variations in 27 natural isolates of the CTG clade yeast *S. stipitis* have demonstrated the variability of its genome. To understand the nature of this variability, the whole genome of the strain NRRL-7124, commonly used in fermentation research, was sequenced by both Illumina and Oxford Nanopore Technologies (ONT) sequencing platforms and compared to the genome of the strain NRRL Y-11545, reference in genomic studies.

Each genome was characterised by the presence of repetitive regions, highlighting the centromeres (previously described), subtelomeres and transposable elements. A total of 6 different families of transposons were identified (3 LTR and 3 non-LTR), of which two, *Ace* and *Can* have, been described for the first time. Overall, the genome of the strain Y-11545 is more repetitive, with the main difference being the number of both transposons and transposon-associated repeats.

The comparison of the whole genome sequence of both strains allowed the detection of a reciprocal translocation between chromosome 5 and 7 as the principle reason behind the karyotype difference observed. The translocation seems associated to instability generated by the presence of repetitive elements in the region (both transposon and non-transposon related). The instability of the genome was further confirmed through adaptive laboratory evolution (ALE) experiments, in which the strain Y-7124 was maintained in exponential growth for two months under temperature stress.

Phenotypic variations have been previously associated with genome modifications in yeasts. This study identified 8 different genome structures in the 27 natural isolates studied, but to our knowledge, no phenotypic studies of most of the strains had been conducted. Hence, the growth parameters of all the natural isolates were analysed in synthetic media with different sugars as carbon source: glucose, xylose, or a mixture of both. Moreover, the effects of inhibitors commonly found in lignocellulose hydrolysates on the growth parameters was studied. Finally, the sedimentation and agar invasion profiles of all the natural isolates was determined, as an indication of their potential to form biofilms. This evaluation demonstrated the correlation between phenotypes and genome structures, with strains belonging to the karyotype group G5 exhibiting the best overall results.

Despite of the relation between repetitive regions in the genome and plasticity, the evolutionary history of the two natural isolates used as reference in this study is unknown. Hence, establishing a direct relationship between the genomic and phenotypic differences observed is complicated. Therefore, the next step was the genome sequencing of 2 strains with improved fermentation phenotypes, adapted from the strain

NRRL Y-7124 following growth in two industrially promising medias for bioethanol production. Although the strain Y-50861 did not show any important differences related to repetitive regions of DNA in coding regions, an aneuploidy was detected in the strain Y-50859. This aneuploidy consisted of the presence of an extra isochromosome (i-5L), stable thanks to the presence of repetitive elements, the telomeres and a centromere, which flanked two identical arms formed by repetitions of a region in chromosome 5. Moreover, the disruption of the *YSA1* gene by a reciprocal translocation non associated with repetitions was detected between chromosome 1 and 2 in the strain Y-50859.

Although no phenotypic study was conducted to correlate the effect of the genome modifications detected between the parental and evolved strains with the improvements in fermentations previously described, several genes have been proposed as candidates to explain these variations.

In conclusion, this project has demonstrated that:

i. Genome plasticity is also observed in CTG clade yeasts other than *Candida* species, and this plasticity is correlated with differences in phenotypes with potential interest for ethanol fermentation.

ii. The genome of *S. stipitis* is rich in repetitive regions and transposable elements, and these condition genome modifications in both long- and short-term evolution.

iii. Genome instability related modifications are promoted during growth in stressful conditions in *S. stipitis*.

iv. Genome conformation conditions phenotypic traits of potential industrial interest in *S. stipitis*. The strains most commonly used in research (Y-11545 for genetic engineering related research and Y-7124 for fermentation studies) are not the best strains according to our phenotypic characterization.

v. The combination of both Oxford Nanopore Technology (ONT) and Illumina sequencing (sequencing of long and short fragments of DNA, respectively) is useful to build assemblies of genomes rich in repetitions and for the quick and effective detection of genome rearrangements and aneuploidies.

References

A Demirci, K-LG Ho and AL Pometto III (1997). Ethanol production by Saccharomyces cerevisiae in biofilm reactors. *Journal of Industrial Microbiology & Biotechnology* **19**:299–304.

Acevedo, A., Conejeros, R. and Aroca, G. (2017). Ethanol production improvement driven by genome-scale metabolic modeling and sensitivity analysis in Scheffersomyces stipitis Yang, S. ed. *PLOS ONE* **12**:e0180074.

Achaz, G. *et al.* (2000). Analysis of Intrachromosomal Duplications in Yeast Saccharomyces cerevisiae: A Possible Model for Their Origin. *Molecular Biology and Evolution* **17**:1268–1275.

Adams, J. *et al.* (1992). Adaptation and major chromosomal changes in populations of Saccharomyces cerevisiae. *Current Genetics* **22**:13–19.

Adams, M.D. (2000). The Genome Sequence of Drosophila melanogaster. *Science* **287**:2185–2195.

Aditiya, H.B. *et al.* (2016). Second generation bioethanol production: A critical review. *Renewable and Sustainable Energy Reviews* **66**:631–653.

Agbogbo, F.K. *et al.* (2006). Fermentation of glucose/xylose mixtures using Pichia stipitis. *Process Biochemistry* **41**:2333–2336.

Agbogbo, F.K. and Coward-Kelly, G. (2008). Cellulosic ethanol production using the naturally occurring xylose-fermenting yeast, Pichia stipitis. *Biotechnology Letters* **30**:1515–1524.

Agbogbo, F.K. and Wenger, K.S. (2006). Effect of pretreatment chemicals on xylose fermentation by Pichia stipitis. *Biotechnology Letters* **28**:2065–2069.

Aguilera, A. and García-Muse, T. (2013). Causes of Genome Instability. *Annual Review of Genetics* **47**:1–32.

Aguilera, A. and García-Muse, T. (2012). R loops: from transcription byproducts to threats to genome stability. *Molecular Cell* **46**:115–124.

Aguilera, A. and Gómez-González, B. (2008). Genome instability: a mechanistic view of its causes and consequences. *Nature Reviews Genetics* **9**:204–217.

Ahmad, K.M. *et al.* (2013). Small chromosomes among Danish Candida glabrata isolates originated through different mechanisms. *Antonie Van Leeuwenhoek* **104**:111–122.

Alba-Lois, L. and Segal-Kischinevzky, C. (2010). Beer & Wine Makers. *Nature Education* **3**:17.

Alberts, B. *et al.* (2002). From DNA to RNA. *Molecular Biology of the Cell. 4th edition* [Online]. Available at: https://www.ncbi.nlm.nih.gov/books/NBK26887/ [Accessed: 4 September 2020].

Aleksenko, A., Nielsen, M.L. and Clutterbuck, A.J. (2001). Genetic and physical mapping of two centromere-proximal regions of chromosome IV in Aspergillus nidulans. *Fungal genetics and biology: FG & B* **32**:45–54.

Alic, N., Higgins, V.J. and Dawes, I.W. (2001). Identification of a Saccharomyces cerevisiae gene that is required for G1 arrest in response to the lipid oxidation

product linoleic acid hydroperoxide. *Molecular Biology of the Cell* **12**:1801–1810.

Alonge, M. *et al.* (2020). Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. *Cell* **182**:145-161.e23.

Alsteens, D. *et al.* (2013). Quantifying the forces driving cell-cell adhesion in a fungal pathogen. *Langmuir : the ACS journal of surfaces and colloids* [Online] **29**. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3858841/ [Accessed: 20 August 2020].

Amartey, S. and Jeffries, T. (1996). An improvement in Pichia stipitis fermentation of acid-hydrolysed hemicellulose achieved by overliming (calcium hydroxide treatment) and strain adaptation. *World Journal of Microbiology & Biotechnology* **12**:281–283.

Amin, G., Standaert, P. and Verachtert, H. (1984). Effects of metabolic inhibitors on the alcoholic fermentation by several yeasts in batch or in immobilized cell systems. *Applied Microbiology and Biotechnology* **19**:91–99.

Andaluz, E., Ciudad, T. and Larriba, G. (2002). An evaluation of the role of LIG4 in genomic instability and adaptive mutagenesis in Candida albicans. *FEMS yeast research* **2**:341–348.

Andersen, E.C. *et al.* (2012). Chromosome-scale selective sweeps shape Caenorhabditis elegans genomic diversity. *Nature Genetics* **44**:285–290.

Andersen, M.P. *et al.* (2008). A genetic screen for increased loss of heterozygosity in Saccharomyces cerevisiae. *Genetics* **179**:1179–1195.

Anderson, J.B. *et al.* (2010). Determinants of Divergent Adaptation and Dobzhansky-Muller Interaction in Experimental Yeast Populations. *Current Biology* **20**:1383–1388.

Antinori, S. *et al.* (2016). Candidemia and invasive candidiasis in adults: A narrative review. *European Journal of Internal Medicine* **34**:21–28.

Aoki, S. and Ito-Kuwa, S. (1984). The appearance and characterization of cyanide-resistant respiration in the fungus Candida albicans. *Microbiology and Immunology* **28**:393–406.

Argueso, J.L. *et al.* (2008). Double-strand breaks associated with repetitive DNA can reshape the genome. *Proceedings of the National Academy of Sciences* **105**:11845–11850.

Arora, R. *et al.* (2014). RNaseH1 regulates TERRA-telomeric DNA hybrids and telomere maintenance in ALT tumour cells. *Nature Communications* **5**:5220.

Atsumi, S., Hanai, T. and Liao, J.C. (2008). Non-fermentative pathways for synthesis of branched-chain higher alcohols as biofuels. *Nature* **451**:86–89.

Augusto Silva De Souza *et al.* (2016). Inhibitors influence on ethanol fermentation by pichia stipitis. *Chemical Engineering Transactions* **49**:367–372.

Avery, O.T., Macleod, C.M. and McCarty, M. (1944). STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES : INDUCTION OF TRANSFORMATION BY A

DESOXYRIBONUCLEIC ACID FRACTION ISOLATED FROM PNEUMOCOCCUS TYPE III. *The Journal of Experimental Medicine* **79**:137–158.

Bajwa, P.K. *et al.* (2011). Ethanol production from selected lignocellulosic hydrolysates by genome shuffled strains of Scheffersomyces stipitis. *Bioresource Technology* **102**:9965–9969.

Bajwa, P.K. *et al.* (2009). Mutants of the pentose-fermenting yeast Pichia stipitis with improved tolerance to inhibitors in hardwood spent sulfite liquor. *Biotechnology and Bioengineering* **104**:892–900.

Bajwa, P.K., Pinel, D., Martin, Vincent J.J., *et al.* (2010). Strain improvement of the pentose-fermenting yeast Pichia stipitis by genome shuffling. *Journal of Microbiological Methods* **81**:179–186.

Bajwa, P.K., Pinel, D., Martin, Vincent J. J., *et al.* (2010). Strain improvement of the pentose-fermenting yeast Pichia stipitis by genome shuffling. *Journal of Microbiological Methods* **81**:179–186.

Balagurunathan, B. *et al.* (2012). Reconstruction and analysis of a genome-scale metabolic model for Scheffersomyces stipitis. *Microbial Cell Factories* **11**:27.

Banat, I.M. *et al.* (1998). Review: ethanol production at elevated temperatures and alcohol concentrations: part I- yeasts in general. *World Journal of Microbiology and Biotechnology* **14**:809–821.

Barrientos, A. (2002). Shy1p is necessary for full expression of mitochondrial COX1 in the yeast model of Leigh's syndrome. *The EMBO Journal* **21**:43–52.

Beckner, M., Ivey, M.L. and Phister, T.G. (2011). Microbial contamination of fuel ethanol fermentations: Bioethanol contamination. *Letters in Applied Microbiology* **53**:387–394.

Belfort, M., Curcio, M.J. and Lue, N.F. (2011). Telomerase and retrotransposons: reverse transcriptases that shaped genomes. *Proceedings of the National Academy of Sciences of the United States of America* **108**:20304–20310.

Bellido, C. *et al.* (2011). Effect of inhibitors formed during wheat straw pretreatment on ethanol fermentation by Pichia stipitis. *Bioresource Technology* **102**:10868–10874.

Bennett, P.M. (2004). Genome Plasticity. In: Woodford, N. and Johnson, A. P. eds. *Genomics, Proteomics, and Clinical Bacteriology*. Totowa, NJ: Humana Press, pp. 71–113. Available at: http://link.springer.com/10.1385/1-59259-763-7:071 [Accessed: 18 August 2020].

Bensasson, D. *et al.* (2008). Rapid Evolution of Yeast Centromeres in the Absence of Drive. *Genetics* **178**:2161–2167.

Bicho, P.A. *et al.* (1988). Induction of Xylose Reductase and Xylitol Dehydrogenase Activities in Pachysolen tannophilus and Pichia stipitis on Mixed Sugars. *Applied and Environmental Microbiology* **54**:50–54.

Biddick, R. and Young, E.T. (2009). The disorderly study of ordered recruitment. *Yeast (Chichester, England)* **26**:205–220.

Biotechnology of Yeasts and Filamentous Fungi (2017). New York, NY: Springer Berlin Heidelberg.

Blackburn, E.H. (1991). Structure and function of telomeres. *Nature* **350**:569–573.

Blank, H.M. *et al.* (2015). Mitotic entry in the presence of DNA damage is a widespread property of aneuploidy in yeast Cohen-Fix, O. ed. *Molecular Biology of the Cell* **26**:1440–1451.

Bleykasten-Grosshans, C., Friedrich, A. and Schacherer, J. (2013). Genome-wide analysis of intraspecific transposon diversity in yeast. *BMC genomics* **14**:399.

Bleykasten-Grosshans, C. and Neuvéglise, C. (2011). Transposable elements in yeasts. *Comptes Rendus Biologies* **334**:679–686.

Botstein, D., Chervitz, S.A. and Cherry, J.M. (1997). Yeast as a model organism. *Science (New York, N.Y.)* **277**:1259–1260.

Bouchonville, K. *et al.* (2009). Aneuploid Chromosomes Are Highly Unstable during DNA Transformation of Candida albicans. *Eukaryotic Cell* **8**:1554–1566.

Bourque, G. *et al.* (2018). Ten things you should know about transposable elements. *Genome Biology* **19**:199.

Brady, T.L. *et al.* (2008). Retrotransposon Target Site Selection by Imitation of a Cellular Protein. *Molecular and Cellular Biology* **28**:1230–1239.

Bravo Ruiz, G. *et al.* (2019). Rapid and extensive karyotype diversification in haploid clinical Candida auris isolates. *Current Genetics* **65**:1217–1228.

Bridges, C.B. (1936). THE BAR 'GENE' A DUPLICATION. *Science* **83**:210–211.

Brown, C.A., Murray, A.W. and Verstrepen, K.J. (2010). Rapid Expansion and Functional Divergence of Subtelomeric Gene Families in Yeasts. *Current Biology* **20**:895–903.

Brown, W.R.A. *et al.* (2011). A Geographically Diverse Collection of Schizosaccharomyces pombe Isolates Shows Limited Phenotypic Variation but Extensive Karyotypic Diversity. *G3: Genes|Genomes|Genetics* **1**:615–626.

Bui, D.T. *et al.* (2015). A Genetic Incompatibility Accelerates Adaptation in Yeast. *PLoS genetics* **11**:e1005407.

Bureau, T.E. and Wessler, S.R. (1992). Tourist: a large family of small inverted repeat elements frequently associated with maize genes. *The Plant Cell* **4**:1283–1294.

Burndy Library, Mendel, G. and Punnett, R.C. (1866). *Versuche Über Pflanzen-Hybriden /*. [Online]. Brünn : Im Verlage des Vereines,. Available at: http://www.biodiversitylibrary.org/bibliography/61004 [Accessed: 19 May 2020].

Buscaino, A. (2019). Chromatin-Mediated Regulation of Genome Plasticity in Human Fungal Pathogens. *Genes* **10**:855.

Butler, G. *et al.* (2009). Evolution of pathogenicity and sexual reproduction in eight Candida genomes. *Nature* **459**:657–662.

Butler, G. (2010). Fungal Sex and Pathogenesis. *Clinical Microbiology Reviews* **23**:140–159.

Cameron, J.R., Loh, E.Y. and Davis, R.W. (1979). Evidence for transposition of dispersed repetitive DNA families in yeast. *Cell* **16**:739–751.

Cao, J. *et al.* (2011). Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nature Genetics* **43**:956–963.

Cao, M., Gao, M., *et al.* (2017). Centromeric DNA Facilitates Nonconventional Yeast Genetic Engineering. *ACS Synthetic Biology* **6**:1545–1553.

Cao, M. *et al.* (2018). CRISPR-Mediated Genome Editing and Gene Repression in Scheffersomyces stipitis. *Biotechnology Journal* **13**:e1700598.

Cao, M., Seetharam, A.S., *et al.* (2017). Rapid Isolation of Centromeres from Scheffersomyces stipitis. *ACS synthetic biology* **6**:2028–2034.

Carlson, M., Celenza, J.L. and Eng, F.J. (1985). Evolution of the dispersed SUC gene family of Saccharomyces by rearrangements of chromosome telomeres. *Molecular and Cellular Biology* **5**:2894–2902.

Carone, D.M. *et al.* (2009). A new class of retroviral and satellite encoded small RNAs emanates from mammalian centromeres. *Chromosoma* **118**:113–125.

Carreras-Puigvert, J. *et al.* (2017). A comprehensive structural, biochemical and biological profiling of the human NUDIX hydrolase family. *Nature Communications* **8**:1541.

Carreté, L. *et al.* (2018). Patterns of Genomic Variation in the Opportunistic Pathogen Candida glabrata Suggest the Existence of Mating and a Secondary Association with Humans. *Current biology: CB* **28**:15-27.e7.

Carreto, L. *et al.* (2008). Comparative genomics of wild type yeast strains unveils important genome diversity. *BMC Genomics* **9**:524.

Carson, A.R. and Scherer, S.W. (2009). Identifying concerted evolution and gene conversion in mammalian gene pairs lasting over 100 million years. *BMC Evolutionary Biology* **9**:156.

Carvalho, C.M.B. and Lupski, J.R. (2016). Mechanisms underlying structural variant formation in genomic disorders. *Nature Reviews. Genetics* **17**:224–238.

Casola, C., Hucks, D. and Feschotte, C. (2007). Convergent Domestication of pogo-like Transposases into Centromere-Binding Proteins in Fission Yeast and Mammals. *Molecular Biology and Evolution* **25**:29–41.

Casper, A.M. *et al.* (2009). Chromosome Aberrations Resulting From Double-Strand DNA Breaks at a Naturally Occurring Yeast Fragile Site Composed of Inverted Ty Elements Are Independent of Mre11p and Sae2p. *Genetics* **183**:423–439.

Castresana, J. (2007). Topological variation in single-gene phylogenetic trees. *Genome Biology* **8**:216.

Cavalieri, D. *et al.* (2018). Genomic and Phenotypic Variation in Morphogenetic Networks of Two Candida albicans Isolates Subtends Their Different Pathogenic Potential. *Frontiers in Immunology* [Online] **8**. Available at: https://www.frontiersin.org/articles/10.3389/fimmu.2017.01997/full [Accessed: 25 August 2020].

Centola, M. and Carbon, J. (1994). Cloning and characterization of centromeric DNA from Neurospora crassa. *Molecular and Cellular Biology* **14**:1510–1519.

Chan, J.E. and Kolodner, R.D. (2011). A Genetic and Structural Study of Genome Rearrangements Mediated by High Copy Repeat Ty1 Elements Haber, J. E. ed. *PLoS Genetics* **7**:e1002089.

Chang, S.-L. *et al.* (2013). Dynamic large-scale chromosomal rearrangements fuel rapid adaptation in yeast populations. *PLoS genetics* **9**:e1003232.

Chargaff, E. *et al.* (1951). The composition of the deoxyribonucleic acid of salmon sperm. *The Journal of Biological Chemistry* **192**:223–230.

Charles, J.S. and Petes, T.D. (2013). High-Resolution Mapping of Spontaneous Mitotic Recombination Hotspots on the 1.1 Mb Arm of Yeast Chromosome IV. *PLOS Genetics* **9**:e1003434.

Chen, G. *et al.* (2012). Hsp90 Stress Potentiates Rapid Cellular Adaptation through Induction of Aneuploidy. *Nature* **482**:246–250.

Chen, G., Rubinstein, B. and Li, R. (2012). Whole chromosome aneuploidy: Big mutations drive adaptation by phenotypic leap. *BioEssays* **34**:893–900.

Chen, S.-H. *et al.* (2012). Engineering Transaldolase in *Pichia stipitis* to Improve Bioethanol Production. *ACS Chemical Biology* **7**:481–486.

Cheng, K.-C., Demirci, A. and Catchmark, J.M. (2010). Advances in biofilm reactors for production of value-added products. *Applied Microbiology and Biotechnology* **87**:445–456.

Chibana, H. *et al.* (1998). A Physical Map of Chromosome 7 of Candida albicans. *Genetics* **149**:1739.

Chibana, H. (2000). Fine-Resolution Physical Mapping of Genomic Diversity in Candida albicans. *Genome Research* **10**:1865–1877.

Chisholm, G.E. and Cooper, T.G. (1992). Ty insertions upstream and downstream of native DUR1,2 promoter elements generate different patterns of DUR1,2 expression in Saccharomyces cerevisiae. *Journal of Bacteriology* **174**:2548–2559.

Cho, J.Y. and Jeffries, T.W. (1998). Pichia stipitis genes for alcohol dehydrogenase with fermentative and respiratory functions. *Applied and Environmental Microbiology* **64**:1350–1358.

Cho, J.Y. and Jeffries, T.W. (1999). Transcriptional control of ADH genes in the xylose-fermenting yeast Pichia stipitis. *Applied and Environmental Microbiology* **65**:2363–2368.

Chu, G., Vollrath, D. and Davis, R. (1986). Separation of large DNA molecules by contour-clamped homogeneous electric fields. *Science* **234**:1582–1585.

Cifuentes-Rojas, C. and Shippen, D.E. (2012). Telomerase regulation. *Mutation Research* **730**:20–27.

Cobb, J.A. (2005). Replisome instability, fork collapse, and gross chromosomal rearrangements arise synergistically from Mec1 kinase and RecQ helicase mutations. *Genes & Development* **19**:3055–3069.

Codón, A.C., Benítez, T. and Korhola, M. (1997). Chromosomal reorganization during meiosis of Saccharomyces cerevisiae baker's yeasts. *Current Genetics* **32**:247–259.

Colón, M. *et al.* (2011). Saccharomyces cerevisiae Bat1 and Bat2 Aminotransferases Have Functionally Diverged from the Ancestral-Like Kluyveromyces lactis Orthologous Enzyme Butler, G. ed. *PLoS ONE* **6**:e16099.

Colson, I., Delneri, D. and Oliver, S.G. (2004). Effects of reciprocal chromosomal translocations on the fitness of *Saccharomyces cerevisiae*. *EMBO reports* **5**:392–398.

Comai, L. (2005). The advantages and disadvantages of being polyploid. *Nature Reviews Genetics* **6**:836–846.

Conant, G.C. and Wolfe, K.H. (2008). Turning a hobby into a job: how duplicated genes find new functions. *Nature Reviews. Genetics* **9**:938–950.

Correns, C.E. (1900). G. Mendel's Regel über das Verhalten der Nachkommenschaft der Rassenbastarde. In: *Gesammelte Abhandlungen zur Vererbungswissenschaft aus Periodischen Schriften 1899–1924*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 9–18. Available at: http://link.springer.com/10.1007/978-3-642-52587-2_2 [Accessed: 20 May 2020].

Corro-Herrera, V.A. *et al.* (2018). Real-time monitoring of ethanol production during *Pichia stipitis* NRRL Y-7124 alcoholic fermentation using transflection near infrared spectroscopy. *Engineering in Life Sciences* **18**:643–653.

Coughlan, A.Y. *et al.* (2016). Centromeres of the Yeast *Komagataella phaffii* (*Pichia pastoris)* Have a Simple Inverted-Repeat Structure. *Genome Biology and Evolution* **8**:2482–2492.

Coughlan, A.Y. and Wolfe, K.H. (2019). The reported point centromeres of Scheffersomyces stipitis are retrotransposon long terminal repeats. *Yeast (Chichester, England)* **36**:275–283.

Cox, E.C. *et al.* (1990). Electrophoretic karyotype for Dictyostelium discoideum. *Proceedings of the National Academy of Sciences* **87**:8247–8251.

Crebelli, R. *et al.* (1989). A comparative study on ethanol and acetaldehyde as inducers of chromosome malsegregation in Aspergillus nidulans. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **215**:187–195.

Crick, F.H. (1958). On protein synthesis. *Symposia of the Society for Experimental Biology* **12**:138–163.

Csink, A.K. and Henikoff, S. (1998). Something from nothing: the evolution and utility of satellite repeats. *Trends in Genetics* **14**:200–204.

Dahm, R. (2008). Discovering DNA: Friedrich Miescher and the early years of nucleic acid research. *Human Genetics* **122**:565–581.

Daley, J.M. *et al.* (2005). Nonhomologous End Joining in Yeast. *Annual Review of Genetics* **39**:431–451.

D'Amore, T. *et al.* (1988). Osmotic pressure effects and intracellular accumulation of ethanol in yeast during fermentation. *Journal of Industrial Microbiology* **2**:365–372.

Davies, G. and Henrissat, B. (1995). Structures and mechanisms of glycosyl hydrolases. *Structure* **3**:853–859.

De Castro, H.F., Oliveira, S.C. and Furlan, S.A. (2003). Alternative Approach for Utilization of Pentose Stream from Sugarcane Bagasse by an Induced Flocculent Pichia stipitis. *Applied Biochemistry and Biotechnology* **107**:547–556.

De Deken, R.H. (1966). The Crabtree Effect: A Regulatory System in Yeast. *Journal of General Microbiology* **44**:149–156.

De Piccoli, G. *et al.* (2006). Smc5-Smc6 mediate DNA double-strand-break repair by promoting sister-chromatid recombination. *Nature Cell Biology* **8**:1032–1034.

De Smet, C.H. *et al.* (2013). Yeast cells accumulate excess endogenous palmitate in phosphatidylcholine by acyl chain remodeling involving the phospholipase B Plb1p. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* **1831**:1167–1176.

Dean, A.M. and Thornton, J.W. (2007). Mechanistic approaches to the study of evolution: the functional synthesis. *Nature Reviews. Genetics* **8**:675–688.

Deininger, P. and Belancio, V.P. (2016). Detection of LINE-1 RNAs by Northern Blot. In: Garcia-Pérez, J. L. ed. *Transposons and Retrotransposons*. New York, NY: Springer New York, pp. 223–236. Available at: http://link.springer.com/10.1007/978-1-4939-3372-3_15 [Accessed: 12 August 2020].

Delgenes, J. P., Moletta, R. and Navarro, J.M. (1988). Continuous production of ethanol from a glucose, xylose and arabinose mixture by a flocculent strain ofPichia stipitis. *Biotechnology Letters* **10**:725–730.

Delgenes, J.P., Moletta, R. and Navarro, J.M. (1988). The ethanol tolerance of Pichia stipitis Y 7124 grown on a d-xylose, d-glucose and l-arabinose mixture. *Journal of Fermentation Technology* **66**:417–422.

Demeke, M.M. *et al.* (2015). Rapid Evolution of Recombinant Saccharomyces cerevisiae for Xylose Fermentation through Formation of Extra-chromosomal Circular DNA Gresham, D. ed. *PLOS Genetics* **11**:e1005010.

Den Haan, R. and Van Zyl, W.H. (2003). Enhanced xylan degradation and utilisation by Pichia stipitis overproducing fungal xylanolytic enzymes. *Enzyme and Microbial Technology* **33**:620–628.

DeRisi, J.L. (1997). Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale. *Science* **278**:680–686.

Deželak, M. *et al.* (2014). The influence of serial repitching of Saccharomyces pastorianus on its karyotype and protein profile during the fermentation of

gluten-free buckwheat and quinoa wort. *International Journal of Food Microbiology* **185**:93–102.

Dietrich, F.S. *et al.* (2004). The Ashbya gossypii genome as a tool for mapping the ancient Saccharomyces cerevisiae genome. *Science (New York, N.Y.)* **304**:304–307.

van Dijk, E.L. *et al.* (2018). The Third Revolution in Sequencing Technology. *Trends in Genetics* **34**:666–681.

van Dijk, M. *et al.* (2019). Strain-dependent variance in short-term adaptation effects of two xylose-fermenting strains of Saccharomyces cerevisiae. *Bioresource Technology* **292**:121922.

Dinh, T.N. *et al.* (2008). Adaptation of Saccharomyces cerevisiae Cells to High Ethanol Concentration and Changes in Fatty Acid Composition of Membrane and Cell Size Herman, C. ed. *PLoS ONE* **3**:e2623.

Douce, R. and Neuburger, M. (1989). The Uniqueness of Plant Mitochondria. *Annual Review of Plant Physiology and Plant Molecular Biology* **40**:371–414.

Dragosits, M. and Mattanovich, D. (2013). Adaptive laboratory evolution – principles and applications for biotechnology. *Microbial Cell Factories* **12**:64.

Du, J., Li, S. and Zhao, H. (2010). Discovery and characterization of novel d-xylose-specific transporters from Neurospora crassa and Pichia stipitis. *Molecular BioSystems* **6**:2150.

Dujon, B. *et al.* (2004). Genome evolution in yeasts. *Nature* **430**:35–44.

Dunham, Maitreya J. *et al.* (2002). Characteristic genome rearrangements in experimental evolution of Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences of the United States of America* **99**:16144–16149.

Dunham, M. J. *et al.* (2002). Characteristic genome rearrangements in experimental evolution of Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences* **99**:16144–16149.

Dunn and Anderson (2019). To Repeat or Not to Repeat: Repetitive Sequences Regulate Genome Stability in Candida albicans. *Genes* **10**:866.

Dunn, B., Levine, R.P. and Sherlock, G. (2005). Microarray karyotyping of commercial wine yeast strains reveals shared, as well as unique, genomic signatures. *BMC Genomics* **6**:53.

Dunn, C.A. *et al.* (1999). Studies on the ADP-ribose pyrophosphatase subfamily of the nudix hydrolases and tentative identification of trgB, a gene associated with tellurite resistance. *The Journal of Biological Chemistry* **274**:32318–32324.

Durkin, S.G. and Glover, T.W. (2007). Chromosome Fragile Sites. *Annual Review of Genetics* **41**:169–192.

Dykhuizen, D. and Hartl, D.L. (1980). Selective neutrality of 6PGD allozymes in E. coli and the effects of genetic background. *Genetics* **96**:801–817.

Dykhuizen, D.E. and Hartl, D.L. (1983). Selection in chemostats. *Microbiological Reviews* **47**:150–168.

Earnshaw, W. *et al.* (1987). Molecular cloning of cDNA for CENP-B, the major human centromere autoantigen. *Journal of Cell Biology* **104**:817–829.

Eder, M. *et al.* (2020). QTL mapping of modelled metabolic fluxes reveals gene variants impacting yeast central carbon metabolism. *Scientific Reports* **10**:2162.

Eid, J. *et al.* (2009). Real-Time DNA Sequencing from Single Polymerase Molecules. *Science* **323**:133–138.

El Hage, A. *et al.* (2010). Loss of Topoisomerase I leads to R-loop-mediated transcriptional blocks during ribosomal RNA synthesis. *Genes & Development* **24**:1546–1558.

Elbarbary, R.A., Lucas, B.A. and Maquat, L.E. (2016). Retrotransposons as regulators of gene expression. *Science* **351**:aac7247–aac7247.

Ellegren, H. (2004). Microsatellites: simple sequences with complex evolution. *Nature Reviews Genetics* **5**:435–445.

Fedoroff, N.V. (2012). Transposable Elements, Epigenetics, and Genome Evolution. *Science* **338**:758–767.

Finnegan, D.J. (1989). Eukaryotic transposable elements and genome evolution. *Trends in Genetics* **5**:103–107.

Fisher, M.C. *et al.* (2018). Worldwide emergence of resistance to antifungal drugs challenges human health and food security. *Science* **360**:739–742.

Fitzgerald-Hayes, M., Clarke, L. and Carbon, J. (1982). Nucleotide sequence comparisons and functional analysis of yeast centromere DNAs. *Cell* **29**:235–244.

Flint, J. *et al.* (1995). The detection of subtelomeric chromosomal rearrangements in idiopathic mental retardation. *Nature Genetics* **9**:132–140.

Forche, A. *et al.* (2011). Stress Alters Rates and Types of Loss of Heterozygosity in Candida albicans Boothroyd, J. C. ed. *mBio* **2**:e00129-11.

Foss, E.J. *et al.* (2017). SIR2 suppresses replication gaps and genome instability by balancing replication between repetitive and unique sequences. *Proceedings of the National Academy of Sciences* **114**:552–557.

Fournier, P. *et al.* (1993). Colocalization of centromeric and replicative functions on autonomously replicating sequences isolated from the yeast Yarrowia lipolytica. *Proceedings of the National Academy of Sciences of the United States of America* **90**:4912–4916.

Frank, A.C. and Wolfe, K.H. (2009). Evolutionary Capture of Viral and Plasmid DNA by Yeast Nuclear Chromosomes. *Eukaryotic Cell* **8**:1521–1531.

Freire-Benéitez, V., Price, R.J., *et al.* (2016). Candida albicans repetitive elements display epigenetic diversity and plasticity. *Scientific Reports* **6**:22989.

Freire-Benéitez, V., Gourlay, S., *et al.* (2016). Sir2 regulates stability of repetitive domains differentially in the human fungal pathogen *Candida albicans*. *Nucleic Acids Research*:gkw594.

Galanty, Y. *et al.* (2012). RNF4, a SUMO-targeted ubiquitin E3 ligase, promotes DNA double-strand break repair. *Genes & Development* **26**:1179–1195.

Gallone, B. *et al.* (2016). Domestication and Divergence of Saccharomyces cerevisiae Beer Yeasts. *Cell* **166**:1397-1410.e16.

Gangloff, S. *et al.* (1994). The yeast type I topoisomerase Top3 interacts with Sgs1, a DNA helicase homolog: a potential eukaryotic reverse gyrase. *Molecular and Cellular Biology* **14**:8391–8398.

Gao, M. *et al.* (2017). Innovating a Nonconventional Yeast Platform for Producing Shikimate as the Building Block of High-Value Aromatics. *ACS Synthetic Biology* **6**:29–38.

GArcía-Aparicio, M.P. *et al.* (2006). Effect of Inhibitors Released During Steam-Explosion Pretreatment of Barley Straw on Enzymatic Hydrolysis. *Applied Biochemistry and Biotechnology* **129**:278–288.

García-Ríos, E., López-Malo, M. and Guillamón, J.M. (2014). Global phenotypic and genomic comparison of two Saccharomyces cerevisiae wine strains reveals a novel role of the sulfur assimilation pathway in adaptation at low temperature fermentations. *BMC genomics* **15**:1059.

Geiger, M. *et al.* (2012). Evaluation of UV-C mutagenized Scheffersomyces stipitis strains for ethanol production. *Journal of Laboratory Automation* **17**:417–424.

Geiler-Samerotte, K. A. *et al.* (2011). Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. *Proceedings of the National Academy of Sciences* **108**:680–685.

Geiler-Samerotte, Kerry A. *et al.* (2011). Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. *Proceedings of the National Academy of Sciences of the United States of America* **108**:680–685.

Génolevures Consortium *et al.* (2009). Comparative genomics of protoploid Saccharomycetaceae. *Genome Research* **19**:1696–1709.

Giaever, G. *et al.* (2002). Functional profiling of the Saccharomyces cerevisiae genome. *Nature* **418**:387–391.

Goffeau, A. *et al.* (1996). Life with 6000 genes. *Science (New York, N.Y.)* **274**:546, 563–567.

Gonçalves, F.A., dos Santos, E.S. and de Macedo, G.R. (2015). Alcoholic fermentation of Saccharomyces cerevisiae, Pichia stipitis and Zymomonas mobilis in the presence of inhibitory compounds and seawater. *Journal of Basic Microbiology* **55**:695–708.

González, J. *et al.* (2017). Diversification of Transcriptional Regulation Determines Subfunctionalization of Paralogous Branched Chain Aminotransferases in the Yeast *Saccharomyces cerevisiae*. *Genetics* **207**:975–991.

Goodier, J.L. (2016). Restricting retrotransposons: a review. *Mobile DNA* **7**:16.

Goodwin, T.J.D., Ormandy, J.E. and Poulter, R.T.M. (2001). L1-like non-LTR retrotransposons in the yeast Candida albicans. *Current Genetics* **39**:83–91.

Gore-Lloyd, D. *et al.* (2019). Snf2 controls pulcherriminic acid biosynthesis and antifungal activity of the biocontrol yeast *Metschnikowia pulcherrima*. *Molecular Microbiology* **112**:317–332.

Gorter de Vries, A.R., Pronk, J.T. and Daran, J.-M.G. (2017). Industrial Relevance of Chromosomal Copy Number Variation in Saccharomyces Yeasts Cullen, D. ed. *Applied and Environmental Microbiology* **83**:e03206-16, e03206-16.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5440705/ [Accessed: 19 August 2020].

Grabundzija, I. *et al.* (2016). A Helitron transposon reconstructed from bats reveals a novel mechanism of genome shuffling in eukaryotes. *Nature Communications* **7**:10716.

Gresham, D. *et al.* (2010). Adaptation to diverse nitrogen-limited environments by deletion or extrachromosomal element formation of the GAP1 locus. *Proceedings of the National Academy of Sciences* **107**:18551–18556.

Gresham, D., Dunham, M.J. and Botstein, D. (2008). Comparing whole genomes using DNA microarrays. *Nature Reviews Genetics* **9**:291–302.

Griffiths, A.J.F. ed. (2012). *Introduction to Genetic Analysis*. 10. ed., 2. print., international ed. New York, NY: Freeman [u.a.].

Grootjen, D.R.J. *et al.* (1991). A flocculating strain of Pichia stipitis for the conversion of glucose/xylose mixtures. *Enzyme and Microbial Technology* **13**:734–739.

Guebel, D.V. *et al.* (1992). Influence of the nitrogen source on growth and ethanol production byPichia stipitis NRRL Y-7124. *Biotechnology Letters* **14**:1193–1198.

Guebel, D.V. and Nudel, C.B. (1994). Spectrophotmetric determination of the settling rate distribution of Pichia stipitis aggregates. *Biotechnology Techniques* **8**:529–534.

Guida, A. *et al.* (2011). Using RNA-seq to determine the transcriptional landscape and the hypoxic response of the pathogenic yeast Candida parapsilosis. *BMC Genomics* **12**:628.

Günan Yücel, H. and Aksu, Z. (2015). Ethanol fermentation characteristics of Pichia stipitis yeast from sugar beet pulp hydrolysate: Use of new detoxification methods. *Fuel* **158**:793–799.

Gurevich, A. *et al.* (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**:1072–1075.

Haas, R. *et al.* (2019). Mapping Ethanol Tolerance in Budding Yeast Reveals High Genetic Variation in a Wild Isolate. *Frontiers in Genetics* **10**:998.

Haber, J.E. (2012). Mating-Type Genes and *MAT* Switching in *Saccharomyces cerevisiae*. *Genetics* **191**:33–64.

Halbeisen, R.E. and Gerber, A.P. (2009). Stress-Dependent Coordination of Transcriptome and Translatome in Yeast Bähler, J. ed. *PLoS Biology* **7**:e1000105.

Hall, A.C. *et al.* (2017). Repetitive DNA loci and their modulation by the non-canonical nucleic acid structures R-loops and G-quadruplexes. *Nucleus* **8**:162–181.

Hansen, K.R. *et al.* (2005). Global effects on gene expression in fission yeast by silencing and RNA interference machineries. *Molecular and Cellular Biology* **25**:590–601.

Hanson, P.K. (2018). *Saccharomyces cerevisiae* : A Unicellular Model Genetic Organism of Enduring Importance: *Saccharomyces cerevisiae* : A Unicellular Model Genetic Organism of Enduring Importance. *Current Protocols Essential Laboratory Techniques* **16**:e21.

Hanson, S.J., Byrne, K.P. and Wolfe, K.H. (2014). Mating-type switching by chromosomal inversion in methylotrophic yeasts suggests an origin for the three-locus Saccharomyces cerevisiae system. *Proceedings of the National Academy of Sciences of the United States of America* **111**:E4851-4858.

Hao, X.-C. *et al.* (2013). Comparative proteomic analysis of a new adaptive Pichia Stipitis strain to furfural, a lignocellulosic inhibitory compound. *Biotechnology for Biofuels* **6**:34.

Harner, N.K. *et al.* (2015). Genetic improvement of native xylose-fermenting yeasts for ethanol production. *Journal of Industrial Microbiology & Biotechnology* **42**:1–20.

Harrison, B.D. *et al.* (2014). A Tetraploid Intermediate Precedes Aneuploid Formation in Yeasts Exposed to Fluconazole Lichten, M. ed. *PLoS Biology* **12**:e1001815.

Hass, E.P. and Zappulla, D.C. (2015). The Ku subunit of telomerase binds Sir4 to recruit telomerase to lengthen telomeres in S. cerevisiae. *eLife* **4**:e07750.

Heather, J.M. and Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics* **107**:1–8.

Hedges, D.J. and Deininger, P.L. (2007). Inviting instability: Transposable elements, double-strand breaks, and the maintenance of genome integrity. *Mutation Research* **616**:46–59.

Hemansi *et al.* (2019). Second Generation Bioethanol Production: The State of Art. In: Srivastava, N. et al. eds. *Sustainable Approaches for Biofuels Production Technologies*. Cham: Springer International Publishing, pp. 121–146. Available at: http://link.springer.com/10.1007/978-3-319-94797-6_8 [Accessed: 20 June 2020].

Henry, S.A. *et al.* (1977). Growth and metabolism of inositol-starved Saccharomyces cerevisiae. *Journal of Bacteriology* **130**:472–484.

Heus, J.J. *et al.* (1993). The consensus sequence of Kluyveromyces lactis centromeres shows homology to functional centromeric DNA from Saccharomyces cerevisiae. *Molecular & general genetics: MGG* **236**:355–362.

Hieter, P. *et al.* (1985). Functional selection and analysis of yeast centromeric DNA. *Cell* **42**:913–921.

Hile, S.E. and Eckert, K.A. (2004). Positive correlation between DNA polymerase alpha-primase pausing and mutagenesis within polypyrimidine/polypurine microsatellite sequences. *Journal of Molecular Biology* **335**:745–759.

Hirakawa, M.P. *et al.* (2015). Genetic and phenotypic intra-species variation in Candida albicans. *Genome Research* **25**:413–425.

Hiraoka, Y., Henderson, E. and Blackburn, E.H. (1998). Not so peculiar: fission yeast telomere repeats. *Trends in Biochemical Sciences* **23**:126.

Hittinger, C.T. *et al.* (2010). Remarkably ancient balanced polymorphisms in a multi-locus gene network. *Nature* **464**:54–58.

Hittinger, C.T. (2013). Saccharomyces diversity and evolution: a budding model genus. *Trends in Genetics* **29**:309–317.

Hoang, M.L. *et al.* (2010). Competitive Repair by Naturally Dispersed Repetitive DNA during Non-Allelic Homologous Recombination. *PLOS Genetics* **6**:e1001228.

Hohmann, S., Krantz, M. and Nordlander, B. (2007). Yeast Osmoregulation. In: *Methods in Enzymology*. Elsevier, pp. 29–45. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0076687907280024 [Accessed: 30 May 2020].

Hope, E.A. and Dunham, M.J. (2014). Ploidy-Regulated Variation in Biofilm-Related Phenotypes in Natural Isolates of Saccharomyces cerevisiae. *G3: Genes|Genomes|Genetics* **4**:1773–1786.

Horváth, V., Merenciano, M. and González, J. (2017). Revisiting the Relationship between Transposable Elements and the Eukaryotic Stress Response. *Trends in genetics: TIG* **33**:832–841.

Hose, J. *et al.* (2015). Dosage compensation can buffer copy-number variation in wild yeast Odom, D. T. ed. *eLife* **4**:e05462.

Hou, J. *et al.* (2017). Engineering of Saccharomyces cerevisiae for the efficient co-utilization of glucose and xylose. *FEMS Yeast Research* [Online] **17**. Available at: https://academic.oup.com/femsyr/article/doi/10.1093/femsyr/fox034/3861258 [Accessed: 20 June 2020].

Huang, C.-F. *et al.* (2009). Enhanced ethanol production by fermentation of rice straw hydrolysate without detoxification using a newly adapted strain of Pichia stipitis. *Bioresource Technology* **100**:3914–3920.

Hughes, S.R. *et al.* (2012). Random UV-C mutagenesis of Scheffersomyces (formerly Pichia) stipitis NRRL Y-7124 to improve anaerobic growth on lignocellulosic sugars. *Journal of Industrial Microbiology & Biotechnology* **39**:163–173.

Huh, W.-K. *et al.* (2003). Global analysis of protein localization in budding yeast. *Nature* **425**:686–691.

Hurst, L.D. and Smith, N.G.C. (1998). The evolution of concerted evolution. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **265**:121–127.

Ilmén, M. *et al.* (2007). Efficient Production of l-Lactic Acid from Xylose by Pichia stipitis. *Applied and Environmental Microbiology* **73**:117–123.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature* **409**:860–921.

International Rice Genome Sequencing Project (2005). The map-based sequence of the rice genome. *Nature* **436**:793–800.

Istace, B. *et al.* (2016). De Novo *Assembly and Population Genomic Survey of Natural Yeast Isolates with the Oxford Nanopore MinION Sequencer*. [Online]. Bioinformatics. Available at: http://biorxiv.org/lookup/doi/10.1101/066613 [Accessed: 23 June 2020].

Ivessa, A.S., Zhou, J.Q. and Zakian, V.A. (2000). The Saccharomyces Pif1p DNA helicase and the highly related Rrm3p have opposite effects on replication fork progression in ribosomal DNA. *Cell* **100**:479–489.

Jacques, N. *et al.* (2010). Population Polymorphism of Nuclear Mitochondrial DNA Insertions Reveals Widespread Diploidy Associated with Loss of Heterozygosity in Debaryomyces hansenii. *Eukaryotic Cell* **9**:449–459.

Jaillon, O. *et al.* (2004). Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature* **431**:946–957.

Jain, M. *et al.* (2015). Improved data analysis for the MinION nanopore sequencer. *Nature Methods* **12**:351–356.

Janbon, G. *et al.* (2010). Characterizing the role of RNA silencing components in Cryptococcus neoformans. *Fungal genetics and biology: FG & B* **47**:1070–1080.

Jasmin, J.-N., Dillon, M.M. and Zeyl, C. (2012). The yield of experimental yeast populations declines during selection. *Proceedings of the Royal Society B: Biological Sciences* **279**:4382–4388.

Jeffares, D.C. *et al.* (2015). The Genomic and Phenotypic Diversity of Schizosaccharomyces pombe. *Nature genetics* **47**:235–241.

Jeffreys, A.J. *et al.* (1988). Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA. *Nature* **332**:278–281.

Jeffries, T.W. (2006). Engineering yeasts for xylose metabolism. *Current Opinion in Biotechnology* **17**:320–326.

Jeffries, T.W. *et al.* (2007). Genome sequence of the lignocellulose-bioconverting and xylose-fermenting yeast Pichia stipitis. *Nature Biotechnology* **25**:319–326.

Jeffries, T.W. and Van Vleet, J.R.H. (2009). *Pichia stipitis* genomics, transcriptomics, and gene clusters. *FEMS Yeast Research* **9**:793–807.

Jeppsson, H., Alexander, N.J. and Hahn-Hagerdal, B. (1995). Existence of Cyanide-Insensitive Respiration in the Yeast Pichia stipitis and Its Possible Influence on Product Formation during Xylose Utilization. *Applied and Environmental Microbiology* **61**:2596–2600.

Jetti, K.D. *et al.* (2019). Improved ethanol productivity and ethanol tolerance through genome shuffling of Saccharomyces cerevisiae and Pichia stipitis. *International Microbiology: The Official Journal of the Spanish Society for Microbiology* **22**:247–254.

Jiao, Y. *et al.* (2011). Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**:97–100.

Jin, Y.-S. *et al.* (2002). Molecular cloning of XYL3 (D-xylulokinase) from Pichia stipitis and characterization of its physiological function. *Applied and Environmental Microbiology* **68**:1232–1239.

Jin, Y.-S., Laplaza, J.M. and Jeffries, T.W. (2004). Saccharomyces cerevisiae engineered for xylose metabolism exhibits a respiratory response. *Applied and Environmental Microbiology* **70**:6816–6825.

Johannsen, W. (1909). *Elemente Der Exakten Erblichkeitslehre. Deutsche Wesentlich Erweiterte Ausgabe in Fünfundzwanzig Vorlesungen, von W. Johannsen.* [Online]. Jena,: G. Fischer,. Available at: http://www.biodiversitylibrary.org/bibliography/1060 [Accessed: 20 May 2020].

Johnson, E.A. and Echavarri-Erasun, C. (2011). Yeast Biotechnology. In: *The Yeasts*. Elsevier, pp. 21–44. Available at: https://linkinghub.elsevier.com/retrieve/pii/B9780444521491000033 [Accessed: 17 June 2020].

Jordan, I.K. and McDonald, J.F. (1998). Evidence for the Role of Recombination in the Regulatory Evolution of Saccharomyces cerevisiae Ty Elements. *Journal of Molecular Evolution* **47**:14–20.

Jutta Hagedorn (1990). *Isolierung Und Charakterisierung von Mutanten Im Xylosestoffwechsel Und Entwicklung Eines Transformationssystems Für Die Hefe Pichia Stipitis*. University of Düsseldorf.

Kaizer, H. *et al.* (2015). Regulation of Telomere Length Requires a Conserved N-Terminal Domain of Rif2 in Saccharomyces cerevisiae. *Genetics* **201**:573–586.

Karagöz, P. and Özkan, M. (2014). Ethanol production from wheat straw by Saccharomyces cerevisiae and Scheffersomyces stipitis co-culture in batch and continuous system. *Bioresource Technology* **158**:286–293.

Karathia, H. *et al.* (2011). Saccharomyces cerevisiae as a Model Organism: A Comparative Study de Polavieja, G. ed. *PLoS ONE* **6**:e16015.

Kasavi, C. *et al.* (2012). Evaluation of industrial Saccharomyces cerevisiae strains for ethanol production from biomass. *Biomass and Bioenergy* **45**:230–238.

Kassiotis, G. and Stoye, J.P. (2016). Immune responses to endogenous retroelements: taking the bad with the good. *Nature Reviews Immunology* **16**:207–219.

Kellis, M., Birren, B.W. and Lander, E.S. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature* **428**:617–624.

Kilian, S.G. and van Uden, N. (1988). Transport of xylose and glucose in the xylose-fermenting yeast Pichia stipitis. *Applied Microbiology and Biotechnology* **27**:545–548.

Kim, J.M. *et al.* (1998). Transposable Elements and Genome Organization: A Comprehensive Survey of Retrotransposons Revealed by the Complete *Saccharomyces cerevisiae* Genome Sequence. *Genome Research* **8**:464–478.

Kim, N. and Jinks-Robertson, S. (2012). Transcription as a source of genome instability. *Nature Reviews. Genetics* **13**:204–214.

Kim, S.R. *et al.* (2012). Simultaneous co-fermentation of mixed sugars: a promising strategy for producing cellulosic ethanol. *Trends in Biotechnology* **30**:274–282.

Kimura, M. (1991). The neutral theory of molecular evolution: a review of recent evidence. *Idengaku Zasshi* **66**:367–386.

Kishida, M. *et al.* (1996). Chromosomal deletion or rearrangement in chimeric hybrids of Saccharomycopsis fibuligera and Saccharomyces diastaticus obtained by cell fusion. *Journal of Fermentation and Bioengineering* **81**:281–285.

Kitada, K. *et al.* (1997). Structural analysis of a Candida glabrata centromere and its functional homology to the Saccharomyces cerevisiae centromere. *Current Genetics* **31**:122–127.

Klimacek, M. *et al.* (2010). Limitations in Xylose-Fermenting Saccharomyces cerevisiae, Made Evident through Comprehensive Metabolite Profiling and Thermodynamic Analysis. *Applied and Environmental Microbiology* **76**:7566–7574.

Klinke, H.B., Thomsen, A.B. and Ahring, B.K. (2004). Inhibition of ethanol-producing yeast and bacteria by degradation products produced during pre-treatment of biomass. *Applied Microbiology and Biotechnology* **66**:10–26.

Klinner, U. *et al.* (2005). Aerobic induction of respiro-fermentative growth by decreasing oxygen tensions in the respiratory yeast Pichia stipitis. *Applied Microbiology and Biotechnology* **67**:247–253.

Ko, B.S., Kim, J. and Kim, J.H. (2006). Production of xylitol from D-xylose by a xylitol dehydrogenase gene-disrupted mutant of Candida tropicalis. *Applied and Environmental Microbiology* **72**:4207–4213.

Ko, J.K. *et al.* (2015). Compounds inhibiting the bioconversion of hydrothermally pretreated lignocellulose. *Applied Microbiology and Biotechnology* **99**:4201–4212.

Koonin, E.V. (2015). Why the Central Dogma: on the nature of the great biological exclusion principle. *Biology Direct* **10**:52.

Koren, S. *et al.* (2017). Canu: scalable and accurate long-read assembly via adaptive *k* -mer weighting and repeat separation. *Genome Research* **27**:722–736.

Kossel A. (1881). *Untersuchungen Über Die Nucleine Und Ihre Spaltungsprodukte*.

Koszul, R., Dujon, B. and Fischer, G. (2006). Stability of Large Segmental Duplications in the Yeast Genome. *Genetics* **172**:2211–2222.

Koti, S. *et al.* (2016). Enhanced bioethanol production from wheat straw hemicellulose by mutant strains of pentose fermenting organisms Pichia stipitis and Candida shehatae. *SpringerPlus* **5**:1545.

Krassowski, T. *et al.* (2018). Evolutionary instability of CUG-Leu in the genetic code of budding yeasts. *Nature Communications* **9**:1887.

Krastanova, O., Hadzhitodorov, M. and Pesheva, M. (2005). Ty Elements of the Yeast *Saccharomyces Cerevisiae*. *Biotechnology & Biotechnological Equipment* **19**:19–26.

Križanović, S. and Butorac, A. (2015). Characterization of a S-adenosyl-l-methionine (SAM)-accumulating strain of Scheffersomyces stipitis. *International Microbiology*:117–125.

Krzywinski, M. *et al.* (2009). Circos: An information aesthetic for comparative genomics. *Genome Research* **19**:1639–1645.

Kumar, S. *et al.* (2018). MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Molecular Biology and Evolution* **35**:1547–1549.

Kumari, R. and Pramanik, K. (2012). Improved bioethanol production using fusants of Saccharomyces cerevisiae and xylose-fermenting yeasts. *Applied Biochemistry and Biotechnology* **167**:873–884.

Kurtz, S. *et al.* (2004). Versatile and open software for comparing large genomes. *Genome Biology* **5**:R12.

Kwak, S. and Jin, Y.-S. (2017). Production of fuels and chemicals from xylose by engineered Saccharomyces cerevisiae: a review and perspective. *Microbial Cell Factories* **16**:82.

Lafontaine, I. and Dujon, B. (2010). Origin and fate of pseudogenes in Hemiascomycetes: a comparative analysis. *BMC Genomics* **11**:260.

Lai, J. *et al.* (2005). Gene movement by Helitron transposons contributes to the haplotype variability of maize. *Proceedings of the National Academy of Sciences* **102**:9068–9073.

Lambers, H. (1982). Cyanide-resistant respiration: A non-phosphorylating electron transport pathway acting as an energy overflow. *Physiologia Plantarum* **55**:478–485.

Lambowitz, A.M. and Slayman, C.W. (1971). Cyanide-resistant respiration in Neurospora crassa. *Journal of Bacteriology* **108**:1087–1096.

Lampe, D.J. *et al.* (1999). Hyperactive transposase mutants of the Himar1 mariner transposon. *Proceedings of the National Academy of Sciences* **96**:11428–11433.

Lan, C.-Y. *et al.* (2002). Metabolic specialization associated with phenotypic switching in Candida albicans. *Proceedings of the National Academy of Sciences* **99**:14907–14912.

Lanciano, S. and Mirouze, M. (2018). Transposable elements: all mobile, all different, some stress responsive, some adaptive? *Current Opinion in Genetics & Development* **49**:106–114.

Landaeta, R. *et al.* (2013). Adaptation of a flocculent Saccharomyces cerevisiae strain to lignocellulosic inhibitors by cell recycle batch fermentation. *Applied Energy* **102**:124–130.

Lange, J.-P. (2007). Lignocellulose conversion: an introduction to chemistry, process and economics. *Biofuels, Bioproducts and Biorefining* **1**:39–48.

de Lange, T. (2005). Shelterin: the protein complex that shapes and safeguards human telomeres. *Genes & Development* **19**:2100–2110.

Laplaza, J.M. *et al.* (2006). Sh ble and Cre adapted for functional genomics and metabolic engineering of Pichia stipitis. *Enzyme and Microbial Technology* **38**:741–747.

Large, C.R.L. *et al.* (2020). *Genomic Stability and Adaptation of Beer Brewing Yeasts during Serial Repitching in the Brewery*. [Online]. Evolutionary Biology. Available at: http://biorxiv.org/lookup/doi/10.1101/2020.06.26.166157 [Accessed: 17 August 2020].

Larriba G (2004). Genome instability, recombination, and adaptation in Candida albicans. In: *Pathogenic Fungi: Host Interactions and Emerging Strategies for Control*. UK: Horizon Press, pp. 285–334.

Larsson, S. *et al.* (1999). Comparison of Different Methods for the Detoxification of Lignocellulose Hydrolyzates of Spruce. *Applied Biochemistry and Biotechnology* **77**:91–104.

Lee, H. *et al.* (1986). Utilization of Xylan by Yeasts and Its Conversion to Ethanol by Pichia stipitis Strains. *Applied and Environmental Microbiology* **52**:320–324.

Lee, K.-T. *et al.* (2014). A Nudix Hydrolase Protein, Ysa1, Regulates Oxidative Stress Response and Antifungal Drug Susceptibility in *Cryptococcus neoformans*. *Mycobiology* **42**:52–58.

Lee, Tae-Hee *et al.* (2001). Estimation of Theoretical Yield for Ethanol Production from D-Xylose by Recombinant Saccharomyces cerevisiae Using Metabolic Pathway Synthesis Algorithm. *Journal of Microbiology and Biotechnology* **11**:384–388.

Lee, T.-Y. *et al.* (2000). A parametric study on ethanol production from xylose byPichia stipitis. *Biotechnology and Bioprocess Engineering* **5**:27–31.

Legrand, M. *et al.* (2019). Candida albicans: An Emerging Yeast Model to Study Eukaryotic Genome Plasticity. *Trends in Genetics* **35**:292–307.

Legrand, M. *et al.* (2007). Role of DNA mismatch repair and double-strand break repair in genome stability and antifungal drug resistance in Candida albicans. *Eukaryotic Cell* **6**:2194–2205.

Lemoine, F.J. *et al.* (2005). Chromosomal translocations in yeast induced by low levels of DNA polymerase a model for chromosome fragile sites. *Cell* **120**:587–598.

Letunic, I. and Bork, P. (2018). 20 years of the SMART protein domain annotation resource. *Nucleic Acids Research* **46**:D493–D496.

Levene P.A., (1919). The Structure of Yeast Nucleic Acid: IV. Ammonia Hydrolysis. *Journal of biological chemistry* **40**:415–424.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997 [q-bio]* [Online]. Available at: http://arxiv.org/abs/1303.3997 [Accessed: 29 June 2020].

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences Birol, I. ed. *Bioinformatics* **34**:3094–3100.

Li, H. *et al.* (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**:2078–2079.

Libuda, D.E. and Winston, F. (2006). Amplification of histone genes by circular chromosome formation in Saccharomyces cerevisiae. *Nature* **443**:1003–1007.

Ligthelm, M.E. *et al.* (1988). An investigation of d-{1-13C} xylose metabolism in Pichia stipitis under aerobic and anaerobic conditions. *Applied Microbiology and Biotechnology* **28**:293–296.

Liti, G. (2015). The fascinating and secret wild life of the budding yeast S. cerevisiae. *eLife* **4**:e05835.

Liu, H. and Zhang, J. (2019). Yeast Spontaneous Mutation Rate and Spectrum Vary with Environment. *Current Biology* **29**:1584-1591.e3.

Liu, M. *et al.* (1997). Structural and mechanistic bases for the induction of mitotic chromosomal loss and duplication (`malsegregation') in the yeast Saccharomyces cerevisiae: Relevance to human carcinogenesis and developmental toxicology. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **374**:209–231.

Lodish H., Berk A and Zipursky SL (2000). *Molecular Cell Biology*. 4th ed. New York: W. H. Freeman.

Loftus, B.J. *et al.* (2005). The genome of the basidiomycetous yeast and human pathogen Cryptococcus neoformans. *Science (New York, N.Y.)* **307**:1321–1324.

Lohe, A.R. and Hartl, D.L. (1996). Autoregulation of mariner transposase activity by overproduction and dominant-negative complementation. *Molecular Biology and Evolution* **13**:549–555.

Louis, E.J. (1995). The chromosome ends ofSaccharomyces cerevisiae. *Yeast* **11**:1553–1573.

Louis Pasteur (1860). *Memoire Sur La Fermentation Alcoolique*. Paris: Mallet-Bachelier.

Lu, P. *et al.* (1998). Cloning and disruption of the b-isopropylmalate dehydrogenase gene (LEU2) of Pichia stipitis with URA3 and recovery of the double auxotroph. *Applied Microbiology and Biotechnology* **49**:141–146.

Lu, P., Davis, B.P. and Jeffries, T.W. (1998). Cloning and characterization of two pyruvate decarboxylase genes from Pichia stipitis CBS 6054. *Applied and Environmental Microbiology* **64**:94–97.

Luikart, G. *et al.* (2003). The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews. Genetics* **4**:981–994.

Luo, K., Vega-Palas, M.A. and Grunstein, M. (2002). Rap1-Sir4 binding independent of other Sir, yKu, or histone interactions initiates the assembly of telomeric heterochromatin in yeast. *Genes & Development* **16**:1528–1539.

Luria, S.E. and Delbrück, M. (1943). Mutations of Bacteria from Virus Sensitivity to Virus Resistance. *Genetics* **28**:491–511.

Lynch, D.B. *et al.* (2010). Chromosomal G + C content evolution in yeasts: systematic interspecies differences, and GC-poor troughs at centromeres. *Genome Biology and Evolution* **2**:572–583.

Lynch, M. *et al.* (2008). A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proceedings of the National Academy of Sciences* **105**:9272–9277.

Lynch. M (2007). *The Origins of Genome Architecture.* 1st ed. Sunderland: Sinauer Associates.

Lynch, M. and Conery, J.S. (2000). The evolutionary fate and consequences of duplicate genes. *Science (New York, N.Y.)* **290**:1151–1155.

Maassen, N. *et al.* (2008). Nonhomologous end joining and homologous recombination DNA repair pathways in integration mutagenesis in the xylose-fermenting yeast *Pichia stipitis*. *FEMS Yeast Research* **8**:735–743.

Mackay, T.F.C. *et al.* (2012). The Drosophila melanogaster Genetic Reference Panel. *Nature* **482**:173–178.

Magee, B.B. *et al.* (2008). Extensive chromosome rearrangements distinguish the karyotype of the hypovirulent species Candida dubliniensis from the virulent Candida albicans. *Fungal genetics and biology : FG & B* **45**:338–350.

Maguire, S.L. *et al.* (2013). Comparative Genome Analysis and Gene Finding in Candida Species Using CGOB. *Molecular Biology and Evolution* **30**:1281–1291.

Mahler, G. and Nudel, C. (2000). Effect of magnesium ions on fermentative and respirative functions in Pichia stipitis under oxygen-restricted growth. *Microbiological Research* **155**:31–35.

Majewski, J. (2000). GT Repeats Are Associated with Recombination on Human Chromosome 22. *Genome Research* **10**:1108–1114.

Malik, H.S. and Henikoff, S. (2009). Major Evolutionary Transitions in Centromere Complexity. *Cell* **138**:1067–1082.

Mandelblat, M. *et al.* (2017). Phenotypic and genotypic characteristics of Candida albicans isolates from bloodstream and mucosal infections. *Mycoses* **60**:534–545.

Mangado, A. *et al.* (2018). Evolution of a Yeast With Industrial Background Under Winemaking Conditions Leads to Diploidization and Chromosomal Copy Number Variation. *Frontiers in Microbiology* **9**:1816.

Mans, R., Daran, J.-M.G. and Pronk, J.T. (2018). Under pressure: evolutionary engineering of yeast strains for improved performance in fuels and chemicals production. *Current Opinion in Biotechnology* **50**:47–56.

Marcand, S. *et al.* (2008). Multiple pathways inhibit NHEJ at telomeres. *Genes & Development* **22**:1153–1158.

Massey, S.E. *et al.* (2003). Comparative evolutionary genomics unveils the molecular mechanism of reassignment of the CTG codon in Candida spp. *Genome Research* **13**:544–557.

Mátés, L. *et al.* (2009). Molecular evolution of a novel hyperactive Sleeping Beauty transposase enables robust stable gene transfer in vertebrates. *Nature Genetics* **41**:753–761.

Mauerhofer, L.-M. *et al.* (2019). Methods for quantification of growth and productivity in anaerobic microbiology and biotechnology. *Folia Microbiologica* **64**:321–360.

McClintock, B. (1950). The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences* **36**:344–355.

McEachern, M.J. and Haber, J.E. (2006). Break-Induced Replication and Recombinational Telomere Elongation in Yeast. *Annual Review of Biochemistry* **75**:111–135.

McEachern, M.J. and Hicks, J.B. (1993). Unusually large telomeric repeats in the yeast Candida albicans. *Molecular and Cellular Biology* **13**:551–560.

McIlwain, S.J. *et al.* (2016). Genome Sequence and Analysis of a Stress-Tolerant, Wild-Derived Strain of *Saccharomyces cerevisiae* Used in Biofuels Research. *G3: Genes|Genomes|Genetics* **6**:1757–1766.

McKenna, A. *et al.* (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**:1297–1303.

Melake, T., Passoth, V. and Klinner, U. (1996). Characterization of the genetic system of the xylose-fermenting yeast Pichia stipitis. *Current Microbiology* **33**:237–242.

Melters, D.P. *et al.* (2013). Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biology* **14**:R10.

Merhej, J. *et al.* (2015). Yap7 is a transcriptional repressor of nitric oxide oxidase in yeasts, which arose from neofunctionalization after whole genome duplication: New insights in the regulation of nitric oxidase. *Molecular Microbiology* **96**:951–972.

Metzger, M.H. and Hollenberg, C.P. (1995). Amino Acid Substitutions in the Yeast Pichia Stipitis Xylitol Dehydrogenase Coenzyme-Binding Domain Affect the Coenzyme Specificity. *European Journal of Biochemistry* **228**:50–54.

Mewes, H.W. *et al.* (1997). Overview of the yeast genome. *Nature* **387**:7–65.

Mézard, C., Pompon, D. and Nicolas, A. (1992). Recombination between similar but not identical DNA sequences during yeast transformation occurs within short stretches of identity. *Cell* **70**:659–670.

Michels, C.A. and Needleman, R.B. (1984). The dispersed, repeated family of MAL loci in Saccharomyces spp. *Journal of Bacteriology* **157**:949–952.

Mieczkowski, P.A. *et al.* (2003). Genetic regulation of telomere-telomere fusions in the yeast Saccharomyces cerevisae. *Proceedings of the National Academy of Sciences* **100**:10854–10859.

Miescher F. (1871). Ueber die chemische Zusammensetzung der Eiterzellen. *Medicinisch-chemische Untersuchungen* **4**.

Mildvan, A.S. *et al.* (2005). Structures and mechanisms of Nudix hydrolases. *Archives of Biochemistry and Biophysics* **433**:129–143.

Min, B., Grigoriev, I.V. and Choi, I.-G. (2017). FunGAP: Fungal Genome Annotation Pipeline using evidence-based gene model evaluation. *Bioinformatics (Oxford, England)* **33**:2936–2937.

Minvielle-Sebastia, L. *et al.* (1997). The major yeast poly(A)-binding protein is associated with cleavage factor IA and functions in premessenger RNA 3′-end formation. *Proceedings of the National Academy of Sciences* **94**:7897–7902.

Mirkin, S.M. (2007). Expandable DNA repeats and human disease. *Nature* **447**:932–940.

Mitrovich, Q.M. *et al.* (2007). Computational and experimental approaches double the number of known introns in the pathogenic yeast Candida albicans. *Genome Research* **17**:492–502.

Mohd Azhar, S.H. *et al.* (2017). Yeasts in sustainable bioethanol production: A review. *Biochemistry and Biophysics Reports* **10**:52–61.

Morange, M. (2009). The Central Dogma of molecular biology: A retrospective after fifty years. *Resonance* **14**:236–247.

Morard, M. *et al.* (2019). Aneuploidy and Ethanol Tolerance in Saccharomyces cerevisiae. *Frontiers in Genetics* [Online] **10**. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6379819/ [Accessed: 19 August 2020].

Morgante, M. *et al.* (2005). Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nature Genetics* **37**:997–1002.

Mortimer, R.K. and Johnston, J.R. (1986). Genealogy of principal strains of the yeast genetic stock center. *Genetics* **113**:35–43.

Moyzis, R.K. *et al.* (1988). A highly conserved repetitive DNA sequence, (TTAGGG)n, present at the telomeres of human chromosomes. *Proceedings of the National Academy of Sciences of the United States of America* **85**:6622–6626.

Munoz-Lopez, M. and Garcia-Perez, J. (2010). DNA Transposons: Nature and Applications in Genomics. *Current Genomics* **11**:115–128.

Murakami, C. and Kaeberlein, M. (2009). Quantifying yeast chronological life span by outgrowth of aged cells. *Journal of Visualized Experiments: JoVE*.

Murphy, T.D. and Karpen, G.H. (1998). Centromeres Take Flight: Alpha Satellite and the Quest for the Human Centromere. *Cell* **93**:317–320.

Mussatto, S.I. *et al.* (2012). Sugars metabolism and ethanol production by different yeast strains from coffee industry wastes hydrolysates. *Applied Energy* **92**:763–768.

Myung, K., Chen, C. and Kolodner, R.D. (2001). Multiple pathways cooperate in the suppression of genome instability in Saccharomyces cerevisiae. *Nature* **411**:1073–1076.

Myung, K., Datta, A. and Kolodner, R.D. (2001). Suppression of Spontaneous Chromosomal Rearrangements by S Phase Checkpoint Functions in Saccharomyces cerevisiae. *Cell* **104**:397–408.

Myung, K. and Kolodner, R.D. (2002). Suppression of genome instability by redundant S-phase checkpoint pathways in Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences* **99**:4500–4507.

Nakamura, K. *et al.* (2008). Rad51 suppresses gross chromosomal rearrangement at centromere in Schizosaccharomyces pombe. *The EMBO journal* **27**:3036–3046.

Narayanan, V. *et al.* (2016). Adaptation to low pH and lignocellulosic inhibitors resulting in ethanolic fermentation and growth of Saccharomyces cerevisiae. *AMB Express* **6**:59.

Naseeb, S. and Delneri, D. (2012). Impact of Chromosomal Inversions on the Yeast DAL Cluster Mata, J. ed. *PLoS ONE* **7**:e42022.

Nasheuer, H.-P. *et al.* (2002). Initiation of eukaryotic DNA replication: regulation and mechanisms. *Progress in Nucleic Acid Research and Molecular Biology* **72**:41–94.

Naumov, G. *et al.* (1990). A new family of polymorphic genes in Saccharomyces cerevisiae: alpha-galactosidase genes MEL1-MEL7. *Molecular & general genetics: MGG* **224**:119–128.

Nei, M. (1969). Gene Duplication and Nucleotide Substitution in Evolution. *Nature* **221**:40–42.

Nei, M., Rogozin, I.B. and Piontkivska, H. (2000). Purifying selection and birth-and-death evolution in the ubiquitin gene family. *Proceedings of the National Academy of Sciences of the United States of America* **97**:10866–10871.

Neuvéglise, C. *et al.* (2005). Mutator-Like Element in the Yeast Yarrowia lipolytica Displays Multiple Alternative Splicings. *Eukaryotic Cell* **4**:615–624.

Nicolaou, S.A., Gaida, S.M. and Papoutsakis, E.T. (2010). A comparative view of metabolite and substrate stress and tolerance in microbial bioprocessing: From biofuels and chemicals, to biocatalysis and bioremediation. *Metabolic Engineering* **12**:307–331.

Niederer, R.O. and Zappulla, D.C. (2015). Refined secondary-structure models of the core of yeast and human telomerase RNAs directed by SHAPE. *RNA (New York, N.Y.)* **21**:1053.

Nigam, J.N. (2001a). Development of xylose-fermenting yeast Pichia stipitis for ethanol production through adaptation on hardwood hemicellulose acid prehydrolysate. *Journal of Applied Microbiology* **90**:208–215.

Nigam, J.N. (2001b). Ethanol production from wheat straw hemicellulose hydrolysate by Pichia stipitis. *Journal of Biotechnology* **87**:17–27.

Nigam, P.S. and Singh, A. (2011). Production of liquid biofuels from renewable resources. *Progress in Energy and Combustion Science* **37**:52–68.

Nikitin, D. *et al.* (2008). Cellular and molecular effects of nonreciprocal chromosome translocations in Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences* **105**:9703–9708.

Nikitin, D.V. *et al.* (2014). Chromosome translocation may lead to PRK1-dependent anticancer drug resistance in yeast via endocytic actin network deregulation. *European Journal of Cell Biology* **93**:145–156.

Nishimura, Y. *et al.* (2018). Metabolic engineering of the 2-ketobutyrate biosynthetic pathway for 1-propanol production in Saccharomyces cerevisiae. *Microbial Cell Factories* **17**:38.

Nogami, S., Ohya, Y. and Yvert, G. (2007). Genetic Complexity and Quantitative Trait Loci Mapping of Yeast Morphological Traits. *PLoS Genetics* **3**:e31.

Novo, M. *et al.* (2009). Eukaryote-to-eukaryote gene transfer events revealed by the genome sequence of the wine yeast Saccharomyces cerevisiae EC1118. *Proceedings of the National Academy of Sciences* **106**:16333–16338.

Nowak, M.A. *et al.* (1997). Evolution of genetic redundancy. *Nature* **388**:167–171.

Nozawa, M. *et al.* (2002). A role of Saccharomyces cerevisiae fatty Acid activation protein 4 in palmitoyl-CoA pool for growth in the presence of ethanol. *Journal of Bioscience and Bioengineering* **93**:288–295.

OhÉigeartaigh, S.S. *et al.* (2011). Systematic discovery of unannotated genes in 11 yeast species using a database of orthologous genomic segments. *BMC genomics* **12**:377.

Ohkuma, M. *et al.* (1995). Identification of a centromeric activity in the autonomously replicating TRA region allows improvement of the host-vector system for Candida maltosa. *Molecular & general genetics: MGG* **249**:447–455.

Ohno, S. (1970). *Evolution by Gene Duplication*. [Online]. Berlin, Heidelberg: Springer Berlin Heidelberg. Available at: http://link.springer.com/10.1007/978-3-642-86659-3 [Accessed: 29 April 2020].

Ohno, S. (1967). *Sex Chromosomes and Sex-Linked Genes*. Berlin: Springer Berlin.

Ola, M. *et al.* (2020). Polymorphic centromere locations in the pathogenic yeast *Candida parapsilosis*. *Genome Research* **30**:684–696.

Olovnikov, A.M. (1973). A theory of marginotomy. *Journal of Theoretical Biology* **41**:181–190.

Onaka, A.T. *et al.* (2016). Rad51 and Rad54 promote noncrossover recombination between centromere repeats on the same chromatid to prevent isochromosome formation. *Nucleic Acids Research* **44**:10744–10757.

Oromendia, A.B., Dodgson, S.E. and Amon, A. (2012). Aneuploidy causes proteotoxic stress in yeast. *Genes & Development* **26**:2696–2708.

Osiro, K.O. *et al.* (2019). Exploring the xylose paradox in Saccharomyces cerevisiae through in vivo sugar signalomics of targeted deletants. *Microbial Cell Factories* **18**:88.

Oud, B. *et al.* (2013). Genome duplication and mutations in ACE2 cause multicellular, fast-sedimenting phenotypes in evolved Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences of the United States of America* **110**:E4223-4231.

Owsianowski, E., Walter, D. and Fahrenkrog, B. (2008). Negative regulation of apoptosis in yeast. *Biochimica Et Biophysica Acta* **1783**:1303–1310.

Padmanabhan, S. *et al.* (2008). Rapid evolution of Cse4p-rich centromeric DNA sequences in closely related pathogenic yeasts, Candida albicans and Candida dubliniensis. *Proceedings of the National Academy of Sciences* **105**:19797–19802.

Paek, A.L. *et al.* (2009). Fusion of nearby inverted repeats by a replication-based mechanism leads to formation of dicentric and acentric chromosomes that cause genome instability in budding yeast. *Genes & Development* **23**:2861–2875.

Paeschke, K., Capra, J.A. and Zakian, V.A. (2011). DNA replication through G-quadruplex motifs is promoted by the Saccharomyces cerevisiae Pif1 DNA helicase. *Cell* **145**:678–691.

Paget, C.M., Schwartz, J.-M. and Delneri, D. (2014). Environmental systems biology of cold-tolerant phenotype in Saccharomyces species adapted to grow at different temperatures. *Molecular Ecology* **23**:5241–5257.

Palmqvist, E. and Hahn-Hägerdal, B. (2000). Fermentation of lignocellulosic hydrolysates. II: inhibitors and mechanisms of inhibition. *Bioresource Technology* **74**:25–33.

Papini, M. *et al.* (2012). Scheffersomyces stipitis: a comparative systems biology study with the Crabtree positive yeast Saccharomyces cerevisiae. *Microbial Cell Factories* **11**:136.

Pâques, F., Leung, W.-Y. and Haber, J.E. (1998). Expansions and Contractions in a Tandem Repeat Induced by Double-Strand Break Repair. *Molecular and Cellular Biology* **18**:2045–2054.

Parapouli, M. *et al.* (2020). Saccharomyces cerevisiae and its industrial applications. *AIMS Microbiology* **6**:1–32.

Pardue, M.-L. and DeBaryshe, P.G. (2011). Retrotransposons that maintain chromosome ends. *Proceedings of the National Academy of Sciences* **108**:20317–20324.

Parry, E.M. and Cox, B.S. (1970). The tolerance of aneuploidy in yeast. *Genetical Research* **16**:333–340.

Parry, J.M. *et al.* (1979). Radiation-induced mitotic and meiotic aneuploidy in the yeast Saccharomyces cerevisiae. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **61**:37–55.

Pascon, R.C. and Miller, B.L. (2000). Morphogenesis in Aspergillus nidulans requires Dopey (DopA), a member of a novel family of leucine zipper-like proteins conserved from yeast to humans. *Molecular Microbiology* **36**:1250–1264.

Passoth, V. *et al.* (1992). The electrophoretic banding pattern of the chromosomes of Pichia stipitis and Candida shehatae. *Current Genetics* **22**:429–431.

Passoth, V., Hahn-Hägerdal, B. and Klinner, U. (2003). Freeze Transformation, Plasmid Reisolation and Stability in Pichia stipitis. In: Wolf, K., Breunig, K. and Barth, G. eds. *Non-Conventional Yeasts in Genetics, Biochemistry and Biotechnology*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 253–259. Available at: http://link.springer.com/10.1007/978-3-642-55758-3_40 [Accessed: 15 June 2020].

Pathania, S., Sharma, N. and Handa, S. (2017). Immobilization of co-culture of *Saccharomyces cerevisiae* and *Scheffersomyces stipitis in sodium alginate* for bioethanol production using hydrolysate of apple pomace under separate hydrolysis and fermentation. *Biocatalysis and Biotransformation* **35**:450–459.

Pavelka, N. *et al.* (2010). Aneuploidy confers quantitative proteome changes and phenotypic variation in budding yeast. *Nature* **468**:321–325.

Peter, J. *et al.* (2018). Genome evolution across 1,011 Saccharomyces cerevisiae isolates. *Nature* **556**:339–344.

Peter, J. and Schacherer, J. (2016). Population genomics of yeasts: towards a comprehensive view across a broad evolutionary scale: Insights into yeast population genomics. *Yeast* **33**:73–81.

Petersen, K.M. and Jespersen, L. (2004). Genetic diversity of the species Debaryomyces hansenii and the use of chromosome polymorphism for typing of strains isolated from surface-ripened cheeses. *Journal of Applied Microbiology* **97**:205–213.

Pfosser, M. *et al.* (1995). Evaluation of sensitivity of flow cytometry in detecting aneuploidy in wheat using disomic and ditelosomic wheat-rye addition lines. *Cytometry* **21**:387–393.

Piednoël, M. *et al.* (2011). Eukaryote DIRS1-like retrotransposons: an overview. *BMC Genomics* **12**:621.

Pierrel, F. *et al.* (2008). Coa2 Is an Assembly Factor for Yeast Cytochrome c Oxidase Biogenesis That Facilitates the Maturation of Cox1. *Molecular and Cellular Biology* **28**:4927–4939.

Pignal, M.-C. (1967). Une nouvelle espèce de levure isolée de larves d'insectes : Pichia stipitis. *Bulletin mensuel de la Société linnéenne de Lyon* **36**:163–168.

Piombo, E. *et al.* (2018). Genome Sequence, Assembly and Characterization of Two Metschnikowia fructicola Strains Used as Biocontrol Agents of Postharvest Diseases. *Frontiers in Microbiology* [Online] **9**. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5891927/ [Accessed: 4 September 2020].

Piontkivska, H., Rooney, A.P. and Nei, M. (2002). Purifying selection and birth-and-death evolution in the histone H4 gene family. *Molecular Biology and Evolution* **19**:689–697.

Plíhal, O. *et al.* (2007). Large Propeptides of Fungal β-N-Acetylhexosaminidases Are Novel Enzyme Regulators That Must Be Intracellularly Processed to Control

Activity, Dimerization, and Secretion into the Extracellular Environment. *Biochemistry* **46**:2719–2734.

Poinsot, C. *et al.* (1987). Isolation and characterization of a mutant of Schwanniomyces castellii with altered respiration. *Antonie Van Leeuwenhoek* **53**:65–75.

Poláková, S. *et al.* (2009). Formation of new chromosomes as a virulence mechanism in yeast *Candida glabrata*. *Proceedings of the National Academy of Sciences* **106**:2688–2693.

Pomés, R., Gil, C. and Nombela, C. (1985). Genetic analysis of Candida albicans morphological mutants. *Journal of General Microbiology* **131**:2107–2113.

Portin, P. and Wilkins, A. (2017). The Evolving Definition of the Term 'Gene'. *Genetics* **205**:1353–1364.

Prado, F. and Aguilera, A. (2005). Impairment of replication fork progression mediates RNA polII transcription-associated recombination. *The EMBO journal* **24**:1267–1276.

Prado, F., Cortés-Ledesma, F. and Aguilera, A. (2004). The absence of the yeast chromatin assembly factor Asf1 increases genomic instability and sister chromatid exchange. *EMBO reports* **5**:497–502.

Pray, L (2008). Discovery of DNA Structure and Function: Watson and Crick. *Nature Education* **1**.

du Preez, J.C., Bosch, M. and Prior, B.A. (1986). The fermentation of hexose and pentose sugars by Candida shehatae and Pichia stipitis. *Applied Microbiology and Biotechnology* **23**:228–233.

du Preez, J.C., van Driessel, B. and Prior, B.A. (1989). Ethanol tolerance of Pichia stipitis and Candida shehatae strains in fed-batch cultures at controlled low dissolved oxygen levels. *Applied Microbiology and Biotechnology* **30**:53–58.

du Preez, J.C. and Prior, B.A. (1985). A quantitative screening of some xylose-fermenting yeast isolates. *Biotechnology Letters* **7**:241–246.

Prior BA, Kilian SG and Du Preez JC (1989). Fermentation of D-xylose by the yeasts Candida shehatae and Pichia stipitis: prospects and problems. *Process Biochemistry* **24**:21–32.

Pryde, F.E. and Louis, E.J. (1999). Limitations of silencing at native yeast telomeres. *The EMBO Journal* **18**:2538–2550.

Querol, A. *et al.* (2003). Adaptive evolution of wine yeast. *International Journal of Food Microbiology* **86**:3–10.

Quinlan, A.R. and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**:841–842.

Quirós, M. *et al.* (2013). Metabolic Flux Analysis during the Exponential Growth Phase of Saccharomyces cerevisiae in Wine Fermentations Polymenis, M. ed. *PLoS ONE* **8**:e71909.

Rachidi, N., Barre, P. and Blondin, B. (1999). Multiple Ty-mediated chromosomal translocations lead to karyotype changes in a wine strain of Saccharomyces cerevisiae. *Molecular and General Genetics MGG* **261**:841–850.

Rancati, G. *et al.* (2008). Aneuploidy Underlies Rapid Adaptive Evolution of Yeast Cells Deprived of a Conserved Cytokinesis Motor. *Cell* **135**:879–893.

Rancati, G. and Pavelka, N. (2013). Karyotypic changes as drivers and catalyzers of cellular evolvability: A perspective from non-pathogenic yeasts. *Seminars in Cell & Developmental Biology* **24**:332–338.

Ratcliff, W.C. *et al.* (2015). Origins of multicellular evolvability in snowflake yeast. *Nature Communications* **6**:6102.

Ray, S. *et al.* (2014). G-quadruplex formation in telomeres enhances POT1/TPP1 protection against RPA binding. *Proceedings of the National Academy of Sciences of the United States of America* **111**:2990–2995.

Ricchetti, M., Dujon, B. and Fairhead, C. (2003). Distance from the Chromosome End Determines the Efficiency of Double Strand Break Repair in Subtelomeres of Haploid Yeast. *Journal of Molecular Biology* **328**:847–862.

Richard, G. and Pâques, F. (2000). Mini- and microsatellite expansions: the recombination connection. *EMBO reports* **1**:122–126.

Richard, G.-F. and Dujon, B. (2006). Molecular Evolution of Minisatellites in Hemiascomycetous Yeasts. *Molecular Biology and Evolution* **23**:189–202.

Robak, K. and Balcerek, M. (2018). Review of Second Generation Bioethanol Production from Residual Biomass. *Food Technology and Biotechnology* **56**:174–187.

Robinson, J.T. *et al.* (2017). Variant Review with the Integrative Genomics Viewer. *Cancer Research* **77**:e31–e34.

Rodrigues, M.V. *et al.* (2007). Bifunctional CTP:Inositol-1-Phosphate Cytidylyltransferase/CDP-Inositol:Inositol-1-Phosphate Transferase, the Key Enzyme for Di-myo-Inositol-Phosphate Synthesis in Several (Hyper)thermophiles. *Journal of Bacteriology* **189**:5405–5412.

Rodrigues, R.C.L.B. *et al.* (2007). Fermentation Kinetics for Xylitol Production by a Pichia stipitis d-Xylulokinase Mutant Previously Grown in Spent Sulfite Liquor. In: Adney, W. S. et al. eds. *Biotechnology for Fuels and Chemicals*. Totowa, NJ: Humana Press, pp. 717–727. Available at: http://link.springer.com/10.1007/978-1-60327-526-2_66 [Accessed: 16 June 2020].

Roelants, F. *et al.* (1995). Reactivation of the ATCase domain of the URA2 gene complex: a positive selection method for Ty insertions and chromosomal rearrangements in Saccharomyces cerevisiae. *Molecular & general genetics: MGG* **246**:767–773.

Rosin, L.F. and Mellone, B.G. (2017). Centromeres Drive a Hard Bargain. *Trends in Genetics* **33**:101–117.

Rothberg, J.M. *et al.* (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**:348–352.

Roy, B. and Sanyal, K. (2011). Diversity in Requirement of Genetic and Epigenetic Factors for Centromere Function in Fungi ▿ . *Eukaryotic Cell* **10**:1384–1395.

Rudd, M.K. (2005). The evolutionary dynamics of -satellite. *Genome Research* **16**:88–96.

Ruhela, D. *et al.* (2015). In vivo role of Candida albicans β-hexosaminidase (HEX1) in carbon scavenging. *MicrobiologyOpen* **4**:730–742.

Rustchenko, E. (2007). Chromosome instability in Candida albicans: Chromosome instability in Candida albicans. *FEMS Yeast Research* **7**:2–11.

Rustchenko-Bulgac, E.P. (1991). Variations of Candida albicans electrophoretic karyotypes. *Journal of Bacteriology* **173**:6586–6596.

Sadhu, C. *et al.* (1991). Telomeric and dispersed repeat sequences in Candida yeasts and their use in strain identification. *Journal of Bacteriology* **173**:842–850.

Saha, A. *et al.* (2015). A *trans* -Dominant Form of Gag Restricts Ty1 Retrotransposition and Mediates Copy Number Control Sundquist, W. I. ed. *Journal of Virology* **89**:3922–3938.

Salazar, A.N. *et al.* (2017). Nanopore sequencing enables near-complete de novo assembly of Saccharomyces cerevisiae reference strain CEN.PK113-7D. *FEMS Yeast Research* [Online] **17**. Available at: https://academic.oup.com/femsyr/article/doi/10.1093/femsyr/fox074/4157789 [Accessed: 22 June 2020].

Salvi, J.S. *et al.* (2014). Roles for Pbp1 and caloric restriction in genome and lifespan maintenance via suppression of RNA-DNA hybrids. *Developmental Cell* **30**:177–191.

SanMiguel, P. *et al.* (1996). Nested Retrotransposons in the Intergenic Regions of the Maize Genome. *Science* **274**:765–768.

Santos, M.A.S. *et al.* (2011). The genetic code of the fungal CTG clade. *Comptes Rendus Biologies* **334**:607–611.

Santos, M.A.S. and Tuite, M.F. (1995). The CUG codon is decoded *in vivo* as serine and not leucine in *Candida albicans*. *Nucleic Acids Research* **23**:1481–1486.

Sanyal, K., Baum, M. and Carbon, J. (2004). Centromeric DNA sequences in the pathogenic yeast Candida albicans are all different and unique. *Proceedings of the National Academy of Sciences of the United States of America* **101**:11374–11379.

Sardi, M. *et al.* (2018). Genome-wide association across Saccharomyces cerevisiae strains reveals substantial variation in underlying gene requirements for toxin tolerance Fay, J. C. ed. *PLOS Genetics* **14**:e1007217.

Sato, T.K. *et al.* (2014). Harnessing Genetic Diversity in Saccharomyces cerevisiae for Fermentation of Xylose in Hydrolysates of Alkaline Hydrogen Peroxide-Pretreated Biomass. *Applied and Environmental Microbiology* **80**:540–554.

Schacherer, J. *et al.* (2009). Comprehensive polymorphism survey elucidates population structure of Saccharomyces cerevisiae. *Nature* **458**:342–345.

Schacherer, J. *et al.* (2004). Recovery of a function involving gene duplication by retroposition in Saccharomyces cerevisiae. *Genome Research* **14**:1291–1297.

Schlötterer, C. and Tautz, D. (1992). Slippage synthesis of simple sequence DNA. *Nucleic Acids Research* **20**:211–215.

Schmieder, R. and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**:863–864.

Schneider, G.F. and Dekker, C. (2012). DNA sequencing with nanopores. *Nature Biotechnology* **30**:326–328.

Schwartz, D.C. and Cantor, C.R. (1984). Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. *Cell* **37**:67–75.

Selmecki, A., Forche, A. and Berman, J. (2006). Aneuploidy and isochromosome formation in drug-resistant Candida albicans. *Science (New York, N.Y.)* **313**:367–370.

Selmecki, A., Forche, A. and Berman, J. (2010). Genomic Plasticity of the Human Fungal Pathogen Candida albicans. *Eukaryotic Cell* **9**:991–1008.

Seoighe, C. *et al.* (2000). Prevalence of small inversions in yeast gene order evolution. *Proceedings of the National Academy of Sciences* **97**:14433–14437.

Serrano, R., Kielland-Brandt, M.C. and Fink, G.R. (1986). Yeast plasma membrane ATPase is essential for growth and has homology with (Na+ + K+), K+- and Ca2+-ATPases. *Nature* **319**:689–693.

Servant, G., Pennetier, C. and Lesage, P. (2008). Remodeling Yeast Gene Transcription by Activating the Ty1 Long Terminal Repeat Retrotransposon under Severe Adenine Deficiency. *Molecular and Cellular Biology* **28**:5543–5554.

Shampay, J., Szostak, J.W. and Blackburn, E.H. (1984). DNA sequences of telomeres maintained in yeast. *Nature* **310**:154–157.

Sharp, J.A. *et al.* (2002). Chromatin assembly factor I and Hir proteins contribute to building functional kinetochores in S. cerevisiae. *Genes & Development* **16**:85–100.

Sheltzer, J. M. *et al.* (2011). Aneuploidy Drives Genomic Instability in Yeast. *Science* **333**:1026–1030.

Sheltzer, Jason M. *et al.* (2011). Aneuploidy drives genomic instability in yeast. *Science (New York, N.Y.)* **333**:1026–1030.

Sheltzer, J.M. *et al.* (2012). Transcriptional consequences of aneuploidy. *Proceedings of the National Academy of Sciences* **109**:12644–12649.

Shen, X.-X. *et al.* (2018). Tempo and mode of genome evolution in the budding yeast subphylum. *Cell* **175**:1533-1545.e20.

Shendure, J. *et al.* (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. *Science (New York, N.Y.)* **309**:1728–1732.

Shi, N.-Q. *et al.* (2002). SHAM-sensitive alternative respiration in the xylose-metabolizing yeast Pichia stipitis. *Yeast (Chichester, England)* **19**:1203–1220.

Shi, N.-Q. and Jeffries, T.W. (1998). Anaerobic growth and improved fermentation of Pichia stipitis bearing a URA1 gene from Saccharomyces cerevisiae. *Applied Microbiology and Biotechnology* **50**:339–345.

Shibata, Y. *et al.* (2009). Yeast genome analysis identifies chromosomal translocation, gene conversion events and several sites of Ty element insertion. *Nucleic Acids Research* **37**:6454–6465.

Silva, J.P.A. *et al.* (2011). Ethanol production from xylose by Pichia stipitis NRRL Y-7124 in a stirred tank bioreactor. *Brazilian Journal of Chemical Engineering* **28**:151–156.

Simão, F.A. *et al.* (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**:3210–3212.

Sirr, A. *et al.* (2015). Allelic variation, aneuploidy, and nongenetic mechanisms suppress a monogenic trait in yeast. *Genetics* **199**:247–262.

Skoneczna, A., Kaniak, A. and Skoneczny, M. (2015). Genetic instability in budding and fission yeast—sources and mechanisms Danchin, A. ed. *FEMS Microbiology Reviews* **39**:917–967.

Skoog, K. *et al.* (1992). Ethanol Reassimilation and Ethanol Tolerance in Pichia stipitis CBS 6054 as Studied by 13C Nuclear Magnetic Resonance Spectroscopy. *Applied and Environmental Microbiology* **58**:2552–2558.

Slininger, P.J. *et al.* (2015). Evolved strains of Scheffersomyces stipitis achieving high ethanol productivity on acid- and base-pretreated biomass hydrolyzate at high solids loading. *Biotechnology for Biofuels* **8**:60.

Slininger, P.J. *et al.* (2006). Nitrogen source and mineral optimization enhance D: -xylose conversion to ethanol by the yeast Pichia stipitis NRRL Y-7124. *Applied Microbiology and Biotechnology* **72**:1285–1296.

Slininger, P.J. *et al.* (1990). Optimum ph and temperature conditions for xylose fermentation byPichia stipitis. *Biotechnology and Bioengineering* **35**:727–731.

Slininger, P.J. *et al.* (2011). Repression of xylose-specific enzymes by ethanol in Scheffersomyces (Pichia) stipitis and utility of repitching xylose-grown populations to eliminate diauxic lag. *Biotechnology and Bioengineering* **108**:1801–1815.

Slininger, P.J., Gorsich, S.W. and Liu, Z.L. (2009). Culture nutrition and physiology impact the inhibitor tolerance of the yeast Pichia stipitis NRRL Y-7124. *Biotechnology and Bioengineering* **102**:778–790.

Slutsky, B. *et al.* (1987). 'White-opaque transition': a second high-frequency switching system in Candida albicans. *Journal of Bacteriology* **169**:189–197.

Slutsky, B., Buffo, J. and Soll, D. (1985). High-frequency switching of colony morphology in Candida albicans. *Science* **230**:666–669.

Smith, C.E., Llorente, B. and Symington, L.S. (2007). Template switching during break-induced replication. *Nature* **447**:102–105.

Smith, D. R. *et al.* (2008). Rapid whole-genome mutational profiling using next-generation sequencing technologies. *Genome Research* **18**:1638–1642.

Smith, Douglas R. *et al.* (2008). Rapid whole-genome mutational profiling using next-generation sequencing technologies. *Genome Research* **18**:1638–1642.

Smith, J.S. *et al.* (2011). Rudimentary G-quadruplex-based telomere capping in Saccharomyces cerevisiae. *Nature Structural & Molecular Biology* **18**:478–485.

Sniegowski, P.D., Dombrowski, P.G. and Fingerman, E. (2002). Saccharomyces cerevisiae and Saccharomyces paradoxus coexist in a natural woodland site in North America and display different levels of reproductive isolation from European conspecifics. *FEMS Yeast Research* **1**:299–306.

Soares, E.V. (2011). Flocculation in Saccharomyces cerevisiae: a review: Yeast flocculation: a review. *Journal of Applied Microbiology* **110**:1–18.

Sogo, J.M. (2002). Fork Reversal and ssDNA Accumulation at Stalled Replication Forks Owing to Checkpoint Defects. *Science* **297**:599–602.

Solieri, L., Dakal, T.C. and Bicciato, S. (2014). Quantitative phenotypic analysis of multistress response in Zygosaccharomyces rouxii complex. *FEMS yeast research* **14**:586–600.

Sood, V. and Brickner, J.H. (2017). Genetic and Epigenetic Strategies Potentiate Gal4 Activation to Enhance Fitness in Recently Diverged Yeast Species. *Current biology: CB* **27**:3591-3602.e3.

Sopko, R. *et al.* (2006). Mapping Pathways and Phenotypes by Systematic Gene Overexpression. *Molecular Cell* **21**:319–330.

Spell, R.M. and Jinks-Robertson, S. (2004). Determination of Mitotic Recombination Rates by Fluctuation Analysis in <I>Saccharomyces cerevisiae</I>. In: *Genetic Recombination*. New Jersey: Humana Press, pp. 003–012. Available at: http://link.springer.com/10.1385/1-59259-761-0:003 [Accessed: 23 June 2020].

Stajich, J.E. *et al.* (2010). Insights into evolution of multicellular fungi from the assembled chromosomes of the mushroom Coprinopsis cinerea (Coprinus cinereus). *Proceedings of the National Academy of Sciences of the United States of America* **107**:11889–11894.

Stearns, T. *et al.* (1990). ADP-ribosylation factor is functionally and physically associated with the Golgi complex. *Proceedings of the National Academy of Sciences of the United States of America* **87**:1238–1242.

Steiner, N.C., Hahnenberger, K.M. and Clarke, L. (1993). Centromeres of the fission yeast Schizosaccharomyces pombe are highly variable genetic loci. *Molecular and Cellular Biology* **13**:4578–4587.

Stephens, S.G. (1951). Possible Significance of Duplication in Evolution. In: *Advances in Genetics*. Elsevier, pp. 247–265. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0065266008602370 [Accessed: 29 April 2020].

Stewart, G. (2018). Yeast Flocculation—Sedimentation and Flotation. *Fermentation* **4**:28.

Stewart, J.A. *et al.* (2012). Maintaining the end: Roles of telomere proteins in end-protection, telomere replication and length regulation. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **730**:12–19.

Stillman, B. (2008). DNA polymerases at the replication fork in eukaryotes. *Molecular cell* **30**:259–260.

Strand, M. *et al.* (1993). Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. *Nature* **365**:274–276.

Sturtevant, A.H. (1925). The Effects of Unequal Crossing over at the Bar Locus in Drosophila. *Genetics* **10**:117–147.

Sturtevant, A.H. and Morgan, T.H. (1923). REVERSE MUTATION OF THE BAR GENE CORRELATED WITH CROSSING OVER. *Science (New York, N.Y.)* **57**:746–747.

Suh, S. *et al.* (2003). Wood ingestion by passalid beetles in the presence of xylose-fermenting gut yeasts. *Molecular Ecology* **12**:3137–3145.

Suzuki, T. *et al.* (1982). Variance of ploidy in Candida albicans. *Journal of Bacteriology* **152**:893–896.

Swinnen, I. (2004). Predictive modelling of the microbial lag phase: a review. *International Journal of Food Microbiology* **94**:137–159.

Swinnen, S. *et al.* (2012). Identification of novel causative genes determining the complex trait of high ethanol tolerance in yeast using pooled-segregant whole-genome sequence analysis. *Genome Research* **22**:975–984.

Szostak, J.W. and Wu, R. (1980). Unequal crossing over in the ribosomal DNA of Saccharomyces cerevisiae. *Nature* **284**:426–430.

Talbot, N.J. and Wayman, M. (1989). Increase in ploidy in yeasts as a response to stressing media. *Applied Microbiology and Biotechnology* **32**:167–169.

Tang, W. *et al.* (2011). Friedreich's Ataxia (GAA)n•(TTC)n Repeats Strongly Stimulate Mitotic Crossovers in Saccharomyces cerevisae Pearson, C. E. ed. *PLoS Genetics* **7**:e1001270.

Taniguchi, M. *et al.* (1997). Ethanol production from a mixture of glucose and xylose by co-culture of Pichia stipitis and a respiratory-deficient mutant of Saccharomyces cerevisiae. *Journal of Fermentation and Bioengineering* **83**:364–370.

Teixeira, M.T. *et al.* (2004). Telomere Length Homeostasis Is Achieved via a Switch between Telomerase- Extendible and -Nonextendible States. *Cell* **117**:323–335.

Theisen, A (2008). Microarray-based Comparative Genomic Hybridization (aCGH). *Nature Education* **1**:45.

Thibessard, A. and Leblond, P. (2014). Subtelomere Plasticity in the Bacterium Streptomyces. In: Louis, E. J. and Becker, M. M. eds. *Subtelomeres*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 243–258. Available at: http://link.springer.com/10.1007/978-3-642-41566-1_14 [Accessed: 12 August 2020].

Thompson, S.L., Bakhoum, S.F. and Compton, D.A. (2010). Mechanisms of Chromosomal Instability. *Current Biology* **20**:R285–R295.

Tinline-Purvis, H. *et al.* (2009). Failed gene conversion leads to extensive end processing and chromosomal rearrangements in fission yeast. *The EMBO journal* **28**:3400–3412.

Tishkoff, D.X. *et al.* (1997). A Novel Mutation Avoidance Mechanism Dependent on S. cerevisiae RAD27 Is Distinct from DNA Mismatch Repair. *Cell* **88**:253–263.

Tiukova, I.A. *et al.* (2019). Chromosomal genome assembly of the ethanol production strain CBS 11270 indicates a highly dynamic genome structure in the yeast species Brettanomyces bruxellensis Schacherer, J. ed. *PLOS ONE* **14**:e0215077.

Todd, Robert T *et al.* (2019). Genome plasticity in Candida albicans is driven by long repeat sequences. *eLife* **8**:e45954.

Todd, Robert T. *et al.* (2019). Genome plasticity in Candida albicans is driven by long repeat sequences. *eLife* **8**.

Todeschini, A.-L. *et al.* (2005). Severe adenine starvation activates Ty1 transcription and retrotransposition in Saccharomyces cerevisiae. *Molecular and Cellular Biology* **25**:7459–7472.

Toivari, M.H. *et al.* (2004). Endogenous Xylose Pathway in Saccharomyces cerevisiae. *Applied and Environmental Microbiology* **70**:3681–3686.

Toivola, A. *et al.* (1984). Alcoholic Fermentation of d-Xylose by Yeasts. *Applied and Environmental Microbiology* **47**:1221–1223.

Tong, L., Lee, S. and Denu, J.M. (2009). Hydrolase regulates NAD+ metabolites and modulates cellular redox. *The Journal of Biological Chemistry* **284**:11256–11266.

Topp, C.N., Zhong, C.X. and Dawe, R.K. (2004). Centromere-encoded RNAs are integral components of the maize kinetochore. *Proceedings of the National Academy of Sciences* **101**:15986–15991.

Török, T., Rockhold, D. and King, A.D. (1993). Use of electrophoretic karyotyping and DNA-DNA hybridization in yeast identification. *International Journal of Food Microbiology* **19**:63–80.

Torres, E.M. *et al.* (2007). Effects of Aneuploidy on Cellular Physiology and Cell Division in Haploid Yeast. *Science* **317**:916–924.

Torres, G.A. *et al.* (2011). Organization and evolution of subtelomeric satellite repeats in the potato genome. *G3 (Bethesda, Md.)* **1**:85–92.

Tørresen, O.K. *et al.* (2019). Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic Acids Research* **47**:10994–11006.

Tosato, V. and Bruschi, C. (2015). Per aspera ad astra: When harmful chromosomal translocations become a plus value in genetic evolution. Lessons from Saccharomyces cerevisiae. *Microbial Cell* **2**:363–375.

Tosato, V., Sidari, S. and Bruschi, C.V. (2013). Bridge-induced chromosome translocation in yeast relies upon a Rad54/Rdh54-dependent, Pol32-independent pathway. *PloS One* **8**:e60926.

Tran, A.V. and Chambers, R.P. (1986). Ethanol fermentation of red oak acid prehydrolysate by the yeast Pichia stipitis CBS 5776. *Enzyme and Microbial Technology* **8**:439–444.

Trautwein, M. *et al.* (2004). Arf1p Provides an Unexpected Link between COPI Vesicles and mRNA in Saccharomyces cerevisiae. *Molecular Biology of the Cell* **15**:5021–5037.

Travers, A. and Muskhelishvili, G. (2015). DNA structure and function. *FEBS Journal* **282**:2279–2295.

Treco, D. and Arnheim, N. (1986). The evolutionarily conserved repetitive sequence d(TG.AC)n promotes reciprocal exchange and generates unusual recombinant tetrads during yeast meiosis. *Molecular and Cellular Biology* **6**:3934–3947.

Tschermak., E (1900). Über künstliche Kreuzung bei Pisum sativum. In: 18th ed. Berichte der Deutschen Botanischen Gesellschaft, pp. 232–239.

Tsushima, A. *et al.* (2019). Genomic Plasticity Mediated by Transposable Elements in the Plant Pathogenic Fungus Colletotrichum higginsianum Van De Peer, Y. ed. *Genome Biology and Evolution* **11**:1487–1500.

Umezu, K. *et al.* (2002). Structural analysis of aberrant chromosomes that occur spontaneously in diploid Saccharomyces cerevisiae: retrotransposon Ty1 plays a crucial role in chromosomal rearrangements. *Genetics* **160**:97–110.

Unrean, P. and Khajeeram, S. (2015). Model-based optimization of Scheffersomyces stipitis and Saccharomyces cerevisiae co-culture for efficient lignocellulosic ethanol production. *Bioresources and Bioprocessing* **2**:41.

Urbina, H., Schuster, J. and Blackwell, M. (2013). The gut of Guatemalan passalid beetles: a habitat colonized by cellobiose- and xylose-fermenting yeasts. *Fungal Ecology* **6**:339–355.

Vaser, R. *et al.* (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research* **27**:737–746.

Vassiliadis, D. *et al.* (2020). Adaptation to Industrial Stressors Through Genomic and Transcriptional Plasticity in a Bioethanol Producing Fission Yeast Isolate. *G3: Genes|Genomes|Genetics* **10**:1375–1391.

Vega, L.R., Mateyak, M.K. and Zakian, V.A. (2003). Getting to the end: telomerase access in yeast and humans. *Nature Reviews Molecular Cell Biology* **4**:948–959.

Venkatesh, A. *et al.* (2018). Draft Genome Sequence of a Highly Heterozygous Yeast Strain from the Metschnikowia pulcherrima Subclade, UCD127. *Genome Announcements* **6**:e00550-18, e00550-18.

Verduyn, C. *et al.* (1985). Properties of the NAD(P)H-dependent xylose reductase from the xylose-fermenting yeast Pichia stipitis. *The Biochemical Journal* **226**:669–677.

Vershinin, A.V., Schwarzacher, T. and Heslop-Harrison, J.S. (1995). The large-scale genomic organization of repetitive DNA families at the telomeres of rye chromosomes. *The Plant Cell* **7**:1823–1833.

Viguera, E. (2001). Replication slippage involves DNA polymerase pausing and dissociation. *The EMBO Journal* **20**:2587–2595.

Villa-Carvajal, M., Querol, A. and Belloch, C. (2006). Identification of species in the genus Pichia by restriction of the internal transcribed spacers (ITS1 and ITS2) and the 5.8S ribosomal DNA gene. *Antonie Van Leeuwenhoek* **90**:171–181.

Viola, A.M. *et al.* (1986). The respiratory activities of four Hansenula species. *Antonie van Leeuwenhoek* **52**:295–308.

Voordeckers, K. *et al.* (2015). Adaptation to High Ethanol Reveals Complex Evolutionary Pathways. *PLoS Genetics* [Online] **11**. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4636377/ [Accessed: 19 August 2020].

Vries, Hugo de (1900). *Sur La Loi de Disjonction Des Hybrides*. Vol. 130. Comptes Rendus de 1'Académie des Sciences.

Waghmare, S.K. and Bruschi, C.V. (2005). Differential chromosome control of ploidy in the yeastSaccharomyces cerevisiae. *Yeast* **22**:625–639.

Walker, B.J. *et al.* (2014). Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement Wang, J. ed. *PLoS ONE* **9**:e112963.

Wallau, G.L., Vieira, C. and Loreto, É.L.S. (2018). Genetic exchange in eukaryotes through horizontal transfer: connected by the mobilome. *Mobile DNA* **9**:6.

Wang, S.S. and Zakian, V.A. (1990). Sequencing of Saccharomyces telomeres cloned using T4 DNA polymerase reveals two domains. *Molecular and Cellular Biology* **10**:4415–4419.

Wang, X. *et al.* (2008). Disruption of Rpn4-Induced Proteasome Expression in *Saccharomyces cerevisiae* Reduces Cell Viability Under Stressed Conditions. *Genetics* **180**:1945–1953.

Wang, Y. (2008). Chromosome instability in yeast and its implications to the study of human cancer. *Frontiers in Bioscience: A Journal and Virtual Library* **13**:2091–2102.

Wang, Z. *et al.* (2019). QTL analysis reveals genomic variants linked to high-temperature fermentation performance in the industrial yeast. *Biotechnology for Biofuels* **12**:59.

Warringer, J. *et al.* (2011). Trait Variation in Yeast Is Defined by Population History. *PLOS Genetics* **7**:e1002111.

Watanabe, T. *et al.* (2011). A UV-induced mutant of Pichia stipitis with increased ethanol production from xylose and selection of a spontaneous mutant with increased ethanol tolerance. *Bioresource Technology* **102**:1844–1848.

Waterhouse, A.M. *et al.* (2009). Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**:1189–1191.

Waterhouse, R.M. *et al.* (2013). OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Research* **41**:D358-365.

Watson, J.D. (1972). Origin of Concatemeric T7DNA. *Nature New Biology* **239**:197–201.

Watson, J.D. and Crick, F.H.C. (1953). Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **171**:737–738.

Wei, L. *et al.* (2015). Engineering Scheffersomyces stipitis for fumaric acid production from xylose. *Bioresource Technology* **187**:246–254.

Weierstall, T., Hollenberg, C.P. and Boles, E. (1999). Cloning and characterization of three genes (SUT1-3) encoding glucose transporters of the yeast Pichia stipitis. *Molecular Microbiology* **31**:871–883.

Welch, J.W., Maloney, D.H. and Fogel, S. (1990). Unequal crossing-over and gene conversion at the amplified CUP1 locus of yeast. *Molecular & general genetics: MGG* **222**:304–310.

Wick, R.R. *et al.* (2017). Completing bacterial genome assemblies with multiplex MinION sequencing. *Microbial Genomics* **3**:e000132.

Wick, R.R., Judd, L.M. and Holt, K.E. (2019). Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biology* **20**:129.

Wicker, T. *et al.* (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics* **8**:973–982.

Wohlbach, D.J. *et al.* (2011). Comparative genomics of xylose-fermenting fungi for enhanced biofuel production. *Proceedings of the National Academy of Sciences* **108**:13212–13217.

Wolfe, K.H. (2015). Origin of the Yeast Whole-Genome Duplication. *PLoS biology* **13**:e1002221.

Wolfe, K.H. and Butler, G. (2017). Evolution of Mating in the Saccharomycotina. *Annual Review of Microbiology* **71**:197–214.

Wolfe, K.H. and Shields, D.C. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**:708–713.

Wöstemeyer, J. and Kreibich, A. (2002). Repetitive DNA elements in fungi (Mycota): impact on genomic architecture and evolution. *Current Genetics* **41**:189–198.

Wu, C.-Y. *et al.* (2010). Control of transcription by cell size. *PLoS biology* **8**:e1000523.

Xie, C. and Tammi, M.T. (2009). CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics* **10**:80.

Xu, L. *et al.* (2019). OrthoVenn2: a web server for whole-genome comparison and annotation of orthologous clusters across multiple species. *Nucleic Acids Research* **47**:W52–W58.

Yadav, V. *et al.* (2018). Five pillars of centromeric chromatin in fungal pathogens. *PLoS Pathogens* [Online] **14**. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6107279/ [Accessed: 2 September 2020].

Yang, M., Ohnuki, S. and Ohya, Y. (2014). Unveiling nonessential gene deletions that confer significant morphological phenotypes beyond natural yeast strains. *BMC genomics* **15**:932.

Yang, V.W. *et al.* (1994). High-efficiency transformation of Pichia stipitis based on its URA3 gene and a homologous autonomous replication sequence, ARS2. *Applied and Environmental Microbiology* **60**:4245–4254.

Yang, X. *et al.* (2011). Ethanol production from the enzymatic hydrolysis of non-detoxified steam-exploded corn stalk. *Bioresource Technology* **102**:7840–7844.

Yona, A.H. *et al.* (2012). Chromosomal duplication is a transient evolutionary solution to stress. *Proceedings of the National Academy of Sciences of the United States of America* **109**:21010–21015.

Yu, Q. *et al.* (2013). The P-type ATPase Spf1 is required for endoplasmic reticulum functions and cell wall integrity in Candida albicans. *International Journal of Medical Microbiology* **303**:257–266.

Yue, J.-X. *et al.* (2017). Contrasting evolutionary genome dynamics between domesticated and wild yeasts. *Nature Genetics* **49**:913–924.

Zarin, T. and Moses, A.M. (2014). Insights into molecular evolution from yeast genomics: Insights into molecular evolution from yeast genomics. *Yeast* **31**:233–241.

Zhang, J. (2003). Evolution by gene duplication: an update. *Trends in Ecology & Evolution* **18**:292–298.

Zhang, W. and Geng, A. (2012). Improved ethanol production by a xylose-fermenting recombinant yeast strain constructed through a modified genome shuffling method. *Biotechnology for Biofuels* **5**:46.

Zhao, J. *et al.* (2010). Non-B DNA structure-induced genetic instability and evolution. *Cellular and molecular life sciences: CMLS* **67**:43–62.

Zhao, M. *et al.* (2013). Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics* **14**:S1.

Zhu, J. *et al.* (2012). Karyotypic Determinants of Chromosome Instability in Aneuploid Budding Yeast Sullivan, B. A. ed. *PLoS Genetics* **8**:e1002719.

Zhu, Y.O. *et al.* (2014). Precise estimates of mutation rate and spectrum in yeast. *Proceedings of the National Academy of Sciences* **111**:E2310–E2318.

Zhu, Yuan O., Sherlock, G. and Petrov, D.A. (2016). Whole Genome Analysis of 132 Clinical Saccharomyces cerevisiae Strains Reveals Extensive Ploidy Variation. *bioRxiv*:044958.

Zhu, Yuan O, Sherlock, G.J. and Petrov, D.A. (2016). *Whole Genome Analysis of 132 Clinical Saccharomyces Cerevisiae Strains Reveals Extensive Ploidy Variation*. [Online]. Genomics. Available at: http://biorxiv.org/lookup/doi/10.1101/044958 [Accessed: 23 June 2020].

Zoghlami, A. and Paës, G. (2019). Lignocellulosic Biomass: Understanding Recalcitrance and Predicting Hydrolysis. *Frontiers in Chemistry* **7**:874.

Zupan, J. and Raspor, P. (2008). Quantitative agar-invasion assay. *Journal of Microbiological Methods* **73**:100–104.

WEB LINKS

https://github.com/harrisonlab/pichia

https://github.com/rrwick/Porechop#how-it-works

https://github.com/ruanjue/smartdenovo

https://github.com/isovic/racon

https://github.com/jts/nanopolish/blob/master/README.md

http://www.repeatmasker.org

http://www.snapgene.com

http://samtools.sourceforge.net

http://www.R-project.org/

https://gatk.broadinstitute.org/hc/en-us/articles/360036194592-Getting-started-with-GATK4

Supplementary material

Table S1. Quality assessment of assembled genomes. N50: Total length of the shortest contig at 50% of the genome length (50% of the genome is contained in contigs larger than this value). L50: Smallest number of contigs whose length sum up makes half of the genome size.

	Y-7124	Y-50859	Y-50861
N. of contigs	11	11	8
Total length (Mbp)	15.69	15.64	15.36
Largest contig (Mbp)	2.69	3.98	3.44
N50 (Mbp)	1.88	1.90	1.90
N75 (Mbp)	1.67	1.32	1.77
L50	4	3	3
L75	6	6	5

Table S2. BUSCO analysis that assess the completeness of genomes according to the identification of single copy orthologs. Database used: saccharomycetales_odb9

	Complete	Duplicated	Fragmented	Missing	Total
Y-7124	1683	13	11	17	1711
Y-50859	1678	9	14	19	1711
Y-50861	1685	5	12	14	1711

Table S3. Coordinates of Transposable elements in the *S. stipitis* strain Y-11545. Red colour indicates conservation in the *S. stipitis* natural isolate Y-7124.

TE	Chromosome	Name	Start	End	Length (bp)
LTR, Copia, Ava	1	1.1	57,601	63,628	6,028
LTR, Copia, Bea	1	1.2	245,155	251,286	6,132
Non-LTR, LINE, L1, Can	1	1.3	440,290	433,543	6,748
LTR, Copia, Ava	1	1.4	1,099,123	1,105,155	6,033
LTR, Copia, Bea	1	1.5	1,282,910	1,276,781	6,130
Non-LTR, LINE, L1, Ace	1	1.6	1,666,970	1,660,557	6,414
LTR, Copia, Bea	1	1.7	1,673,100	1,666,969	6,132
Non-LTR, LINE, L1, Ace	1	1.8	1,719,581	1,714,009	5,573
Non-LTR, LINE, L1, Ace	1	1.9	1,780,043	1,786,974	6,932
Non-LTR, LINE, L1, Ace	1	1.10	1,907,951	1,901,045	6,907
LTR, Copia, Bea	1	1.11	2,182,419	2,176,410	6,010
LTR, Copia, Caia	1	1.12	2,236,681	2,243,255	6,575
Non-LTR, LINE, L1, Ace	1	1.13	2,254,338	2,257,330	2,993
N/A	1	1.14	2,547,473	2,559,254	11,782
Non-LTR, LINE, L1, Bri	1	1.15	3,081,925	3,076,138	5,788
Non-LTR, LINE, L1, Bri	2	2.0	387,620	388,452	833
LTR, Copia, Caia	2	2.1	643,160	636,595	6,566
Non-LTR, LINE, L1, Ace	2	2.2	924,508	931,387	6,800
Non-LTR, LINE, L1, Can	2	2.8	931,388	940,017	8,630
LTR, Copia, Ava	2	2.3	1,601,129	1,607,172	6,044
LTR, Copia, Ava	2	2.4	1,981,078	1,987,102	6,025
Non-LTR, LINE, L1, Can	2	2.5	2,112,171	2,118,885	6,715
Non-LTR, LINE, L1, Ace	2	2.7	2,593,318	2,600,238	6,921
Non-LTR, LINE, L1, Bri	2	2.9	2,587,549	2,593,336	5,788
LTR, Copia, Caia	3	3.1	202,368	195,802	6,567
Non-LTR, LINE, L1, Ace	3	3.2	459,509	466,345	6,837
Non-LTR, LINE, L1, Ace	3	3.3	609,371	602,625	6,747
Non-LTR, LINE, L1, Ace	3	3.4	1390766	1390766	7,010
Non-LTR, LINE, L1, Ace	3	3.5	1719797	1713055	6,743
Non-LTR, LINE, L1, Bri	3	3.6	1,622,685	1,628,466	5,782
Non-LTR, LINE, L1, Can	4	4.1	274,056	280,113	6,058
Non-LTR, LINE, L1, Can	4	4.7	280,122	286,657	6,536
LTR, Copia, Bea	4	4.2	1,098,750	1,104,883	6,134
Non-LTR, LINE, L1, Ace	4	4.6	1,135,307	1,142,100	6,794
LTR, Copia, Ava	4	4.3	1,459,635	1,453,602	6,034
LTR, Copia, Bea	4	4.4	1,724,824	1,718,695	6,130
Non-LTR, LINE, L1, Bri	5	5.1	549,428	555,340	5,913
Non-LTR, LINE, L1, Ace	5	5.2	710728	709380	1,349
Non-LTR, LINE, L1, Can	5	5.3	1,377,252	1,370,447	6,806
LTR, Copia, Ava	6	6.1	571,952	565,924	6,029
LTR, copia, N/A	6	6.2	724200	722896	1,305
Non-LTR, LINE, L1, Ace	6	6.3	1,389,654	1,382,694	6,961
LTR, Copia, Bea	7	7.1	456,397	450,433	5,965
LTR, Copia, Caia	7	7.3	1,073,204	1,066,648	6,557
Non-LTR, LINE, L1, Ace	8	8.1	253,230	260,163	6,934
Non-LTR, LINE, L1, Ace	8	8.3	260,174	263,137	2,694
LTR, Copia, Bea	8	8.2	899,027	905,166	6,140

Table S4. Coordinates of Transposable elements in the *S. stipitis* strain Y-7124. Red colour indicates conservation in the *S. stipitis* natural isolate Y-11545.

TE	Chromosome	Name	Start	End	Length (bp)
N/A	1	1.0	2,485,152	2,497,030	11,879
Non-LTR, LINE, L1, Bri	1	1.1	2,443,220	2,448,999	5,780
Non-LTR, LINE, L1, Can	1	1.2	2,193,343	2,196,597	3,255
Non-LTR, LINE, L1, Ace	1	1.3	154,001	160,945	6,945
Non-LTR, LINE, L1, Can	2	2.1	2,217,025	2,221,424	4,400
Non-LTR, LINE, L1, Can	2	2.2	1,304,980	1,311,506	6,527
Non-LTR, LINE, L1, Can	2	2.3	976,356	982,881	6,527
Non-LTR, LINE, L1, Bri	2	2.4	385,643	386,475	833
LTR, Copia, Bea	3	3.1	271,048	264,930	6,119
Non-LTR, LINE, L1, Can	3	3.2	1,380,965	1,374,439	6,527
Non-LTR, LINE, L1, Ace	4	4.1	113,819	108,059	5,761
Non-LTR, LINE, L1, Can	4	4.2	227,369	222,731	4,639
Non-LTR, LINE, L1, Can	4	4.3	439,930	438,767	1,164
LTR, Copia, Ava	4	4.4	494,021	500,049	6,029
Non-LTR, LINE, L1, Ace	4	4.5	1,262,814	1,269,555	6,742
Non-LTR, LINE, L1, Ace	4	4.6	1,274,737	1,281,260	6,524
LTR, Copia, Caia	5	5.1	1,190,258	1,196,818	6,561
Non-LTR, LINE, L1, Ace	5	5.3	561,001	567,259	6,259
Non-LTR, LINE, L1, Ace	5	5.4	879,485	880,621	1,137
Non-LTR, LINE, L1, Bri	5	5.5	568,197	570,428	2,232
N/A	6	6.1	689,104	690,403	1,305
Non-LTR, LINE, L1, Bri	7	7.1	134,602	140,374	5,573

Table S5. Coordinates of Transposable elements in the *S. stipitis* strain Y-50859. All TE elements were conserved when compared to the *S. stipitis* parental strain Y-7124.

TE	Chromosome	Name	Start	End	Length (bp)
N/A	1	1.0	2,484,985	2,496,863	11,879
Non-LTR, LINE, L1, Bri	1	1.1	2,443,049	2,448,819	5,771
Non-LTR, LINE, L1, Can	1	1.2	2,193,169	2,196,424	3,256
Non-LTR, LINE, L1, Ace	1	1.3	153,850	160,795	6,946
Non-LTR, LINE, L1, Can	2	2.1	3,000,076	3,006,601	6,526
Non-LTR, LINE, L1, Can	2	2.2	3,596,484	3,597,316	833
Non-LTR, LINE, L1, Can	2	2.3	1,629,112	1,633,509	4,398
Non-LTR, LINE, L1, Bri	2	2.4	717,006	723,535	6,530
LTR, Copia, Bea	3	3.1	265,184	271,302	6,119
Non-LTR, LINE, L1, Can	3	3.2	1,374,640	1,381,164	6,525
Non-LTR, LINE, L1, Ace	4	4.1	108,329	114,088	5,760
Non-LTR, LINE, L1, Can	4	4.2	223,000	227,639	4,640
Non-LTR, LINE, L1, Can	4	4.3	439,037	440,202	1,166
LTR, Copia, Ava	4	4.4	494,293	500,321	6,029
Non-LTR, LINE, L1, Ace	4	4.5	1,263,085	1,269,829	6,745
Non-LTR, LINE, L1, Ace	4	4.6	1,275,011	1,281,535	6,525
LTR, Copia, Caia	5	5.1	1,194,820	1,201,380	6,561
Non-LTR, LINE, L1, Ace	5	5.3	560,833	567,091	6,259
Non-LTR, LINE, L1, Ace	5	5.4	884,045	885,181	1,137
Non-LTR, LINE, L1, Bri	5	5.5	568,029	570,260	2,232
N/A	6	6.1	724,785	726,084	1,300
Non-LTR, LINE, L1, Bri	7	7.1	134,676	140,452	5,777

Table S6. Coordinates of Transposable elements in the *S. stipitis* strain Y-50861. All TE elements were conserved when compared to the *S. stipitis* parental strain Y-7124.

TE	Chromosome	Name	Start	End	Length (bp)
N/A	1	1.0	2,488,277	2,500,155	11,879
Non-LTR, LINE, L1, Bri	1	1.1	2,446,346	2,452,120	5,575
Non-LTR, LINE, L1, Can	1	1.2	2,193,153	2,196,407	2,355
Non-LTR, LINE, L1, Ace	1	1.3	153,839	160,783	6,945
Non-LTR, LINE, L1, Can	2	2.1	2,213,986	2,218,386	4,401
Non-LTR, LINE, L1, Can	2	2.2	1,303,092	1,309,620	6,529
Non-LTR, LINE, L1, Can	2	2.3	976,339	982,865	6,527
Non-LTR, LINE, L1, Bri	2	2.4	385,626	386,458	833
LTR, Copia, Bea	3	3.1	265,079	271,197	6,119
Non-LTR, LINE, L1, Can	3	3.2	1,374,517	1,381,043	6,527
Non-LTR, LINE, L1, Ace	4	4.1	107,997	113,757	5,761
Non-LTR, LINE, L1, Can	4	4.2	222,669	227,307	4,369
Non-LTR, LINE, L1, Can	4	4.3	438,705	439,868	1,164
LTR, Copia, Ava	4	4.4	493,959	499,987	6,029
Non-LTR, LINE, L1, Ace	4	4.5	1,262,757	1,269,498	6,742
Non-LTR, LINE, L1, Ace	4	4.6	1,274,674	1,280,814	6,141
LTR, Copia, Caia	5	5.1	1,194,530	1,201,090	6,561
Non-LTR, LINE, L1, Ace	5	5.3	560,539	566,797	6,259
Non-LTR, LINE, L1, Ace	5	5.4	883,757	884,893	1,137
Non-LTR, LINE, L1, Bri	5	5.5	567,735	569,966	2,232
N/A	6	6.1	725,113	726,030	1,300
Non-LTR, LINE, L1, Bri	7	7.1	134,593	140,382	5,790