Catarina Marreiros Lagoas

Licenciada em Biologia

A broad evolutionary perspective of alcoholic fermentation in a non-conventional yeast clade

Dissertação para obtenção do Grau de Mestre em
Genética Molecular e Biomedicina

Orientadora: Doutora Carla Gonçalves, PostDoc, FCT/UNL

Co-orientadora: Professora Doutora Paula Gonçalves, Profª Associada, FCT/UNL

Júri

Presidente:	Professora Doutora Margarida Casal Ribeiro Castro Caldas Braga
Arguente:	Professor Doutor Pedro Manuel Brôa Costa
Vogal:	Doutora Carla Isabel Gomes Gonçalves

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Junho, 2021

Catarina Marreiros Lagoas

Licenciada em Biologia

A broad evolutionary perspective of alcoholic fermentation in a non-conventional yeast clade

Dissertação para obtenção do Grau de Mestre em
Genética Molecular e Biomedicina

Orientadora: Doutora Carla Gonçalves, PostDoc, FCT/UNL

Co-orientadora: Professora Doutora Paula Gonçalves, Profª Associada, FCT/UNL

Júri

Presidente:	Professora Doutora Margarida Casal Ribeiro Castro Caldas Braga
Arguente:	Professor Doutor Pedro Manuel Brôa Costa
Vogal:	Doutora Carla Isabel Gomes Gonçalves

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Junho, 2021

A broad evolutionary perspective of alcoholic fermentation in a non-conventional yeast clade

Acknowledgements

I would like to express my gratitude towards my supervisor, Doctor Carla Gonçalves, for all the guidance, encouragement and tireless help during this journey. I really appreciate all the conversations we have had that taught me a lot and always kept me motivated.

I would also like to thank my co-supervisor, Professor Paula Gonçalves, for giving me the opportunity to do my dissertation in the laboratory and for all the help, suggestions and guidance throughout this experience.

A special thanks to all the members of the Yeast Genomics Lab for always making me feel welcomed in the lab and for everything I have learnt from them.

Also acknowledged is Portuguese Yeast Culture Collection (PYCC) that provided all the strains that were necessary for this work and the Y1000+ Project that contributed with a part of the genomes that were analysed.

I would also like to give my appreciation to Filipa Garvão not only for the molecular constructs that were used in this project but also for her constant enthusiasm with the following steps of the experiments. I am also thankful to Nuno Costa from the Associated Laboratory for Green Chemistry that promptly analysed many of the HPLC samples for this project. I would also like to thank Professor Sérgio Filipe and Professor Jaime Mota for providing other necessary materials for the experiments and Doctor Inês Grilo for all the suggestions regarding protein purification.

A very special thanks goes to all my close family and friends whose immeasurable support and care were essential for the conclusion of this dissertation.

Resumo

A evolução da fermentação alcoólica no clado de leveduras não convencionais *Wickerhamiella*/*Starmerella* (W/S) é caracterizada pela perda dos genes nativos da piruvato descarboxilase (*PDC1*) e das álcool desidrogenases (*ADH*). Em algumas espécies, a reaquisição desta via foi conseguida através de transferências horizontais de genes (HGT) *ADH* e pela cooptação de uma descarboxilase nativa, Aro10. Este trabalho teve como objetivo partir do conhecimento prévio relativo à fermentação alcoólica no clado W/S, combinando dados *in silico* de novos genomas sequenciados, com ensaios fenotípicos, de forma a avaliar as capacidades fermentativas e caracterizar as enzimas Adh, num conjunto de espécies. Três eventos HGT independentes foram previamente identificados como tendo introduzido diferentes *ADH1* bacterianos nos distintos subgrupos do clado W/S. Os subgrupos A, B e C possuem *ADH1a*, *ADH1b* e *ADH1c*, respetivamente. O subgrupo *ADH0* não possui *ADH1*. Neste trabalho, dados que suportam estes três eventos HGT foram obtidos e dois novos eventos HGT de *ADH6* bacterianos foram detetados. A maioria dos genes *ADH6* foram adquiridos nos mesmos ancestrais dos subgrupos reportados para *ADH1*, enquanto um foi encontrado numa espécie *ADH0* (*Wickerhamiella slavikovae*) que aparentemente não possui outros genes da fermentação alcoólica (*ADH1*, *PDC1* e *ARO10*). Em relação às espécies restantes, enquanto *ARO10* está presente, *PDC1* está ausente. A produção de etanol foi geralmente observada no subgrupo A, enquanto a sua assimilação foi verificada nos subgrupos B e C, sugerindo que as proteínas Adh são funcionais. Foi confirmado o papel da Adh1a de *Starmerella bombicola* na interconversão de acetaldeído e etanol, usando NAD(H) e NADP(H) como cofatores, o que contrasta com a especificidade das proteínas Adh de leveduras relativamente a NAD(H). Para compreender melhor a evolução da fermentação alcoólica no clado W/S, é essencial combinar genómica comparativa com a caracterização destas enzimas, de forma a avaliar o seu papel no metabolismo central de carbono.

Termos-chave:
Transferência horizontal de genes, fermentação alcoólica, *Wickerhamiella*, *Starmerella*, álcool desidrogenase, piruvato descarboxilase

Abstract

The evolution of alcoholic fermentation in the non-conventional *Wickerhamiella*/*Starmerella* (W/S) yeast clade is characterized by the loss of native pyruvate decarboxylase (*PDC1*) and alcohol dehydrogenase (*ADH*) genes. In some species, the reacquisition of this via was achieved through horizontal gene transfer (HGT) of *ADH* and co-option of a native decarboxylase, Aro10. This work aimed to build on the previous knowledge regarding alcoholic fermentation in the W/S clade, by combining *in silico* data obtained from newly sequenced genomes, with phenotypic assays, aiming to assess alcoholic fermentation abilities and characterization of Adh enzymes, in a subset of species. Three independent HGT events were previously reported to have introduced different bacterial *ADH1* in distinct subgroups of the W/S clade. Subgroups A, B and C harbour *ADH1a*, *ADH1b* and *ADH1c*, respectively. Subgroup *ADH0* does not carry an *ADH1*. In this work, data supporting these three HGT events was obtained and two new HGT events of bacterial *ADH6* were detected. Most *ADH6* genes were acquired in the ancestor of the subgroups reported for *ADH1*, while one was found in a *ADH0* species (*Wickerhamiella slavikovae*) that seemingly lacks other alcoholic fermentation genes (*ADH1*, *PDC1* and *ARO10*). As for the remaining species, while *ARO10* is present, *PDC1* is absent. Ethanol production was generally observed in subgroup A, while its assimilation was verified in subgroups B and C, suggesting that Adh proteins are functional. It was confirmed that Adh1a from *Starmerella bombicola* is involved in the interconversion of acetaldehyde and ethanol, using NAD(H) and NADP(H) as cofactors, which contrasts with the specificity of Adh proteins from yeasts towards NAD(H). To further understand the evolution of the alcoholic fermentation in the W/S clade, it is essential to combine comparative genomics with the characterization of these enzymes to evaluate their role on the central carbon metabolism.

Key words:
Horizontal gene transfer, alcoholic fermentation, *Wickerhamiella*, *Starmerella*, alcohol dehydrogenase, pyruvate decarboxylase

Table of Contents

List of Figures

List of Tables

Abbreviations

APS	Ammonium persulfate
ATP	Adenosine triphosphate
BLAST	Basic local alignment search tool
bp	Base pair(s)
CBS	Centraalbureau voor Schimmelcultures, Utrecht, The Netherlands
CD-HIT	Cluster database at high identity with tolerance
CDS	Coding sequence(s)
DNA	Desoxyribonucleic acid
DTT	Dithiothreitol
Ffz1	Fructose facilitator Zygosaccharomyces 1
FLAB	Fructophilic lactic acid bacteria
g	Acceleration of gravity
h	Hour(s)
HGT	Horizontal gene transfer
HPLC	High-performance liquid chromatography
IB	Inclusion bodies
IMAC	Immobilized metal ion affinity chromatography
IPTG	Isopropyl β- d-1-thiogalactopyranoside
ITS	Internal transcribed spacer
kb	Kilobase(s)
MAFFT	Multiple alignment program for amino acid or nucleotide sequences
min	Minute(s)
ML	Maximum likelihood
MRCA	Most recent common ancestor
mRNA	Messenger ribonucleic acid
NAD(H)	Nicotinamide adenine dinucleotide
NADP(H)	Nicotinamide adenine dinucleotide phosphate
NCBI	National Centre for Biotechnology Information
nm	Nanometres
NRRL	Agriculture Research Service (ARS) Culture Collection, National Centre for Agricultural Utilization Research, Peoria, Illinois, USA
OD	Optical density
PCR	Polymerase chain reaction

PMSF	Phenylmethylsulphonyl fluoride
PYCC	Portuguese Yeast Culture Collection, CREM, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Caparica, Portugal
rRNA	Ribosomal ribonucleic acid
SDS	Sodium dodecyl sulphate
s	Second(s)
SDS-PAGE	SDS–polyacrylamide gel electrophoresis
TEMED	Tetramethylethylenediamine
TGS	Tris-Glycine-SDS buffer
TRIS	Tris(hydroxymethyl)aminomethane
tRNA	Transfer ribonucleic acid
v/v	Volume per volume
W/S	*Wickerhamiella*/*Starmerella*
w/v	Weight per volume
WGD	Whole genome duplication

1. Introduction

1.1. Horizontal Gene Transfer (HGT)

Horizontal Gene Transfer (HGT), also designated as lateral or nonvertical gene transfer, has a crucial role in the evolution of species, providing adaptation to new environments. It consists in the movement of genetic material between two reproductively isolated genomes, overcoming in this way the mating barriers between different organisms. HGT is a widespread mechanism, present across all three domains of life and it can occur between distant or closely related species. Even though HGT is extremely frequent between *Bacteria*, these events are far more uncommon in *Eukarya* (Keeling & Palmer, 2008; Milner et al., 2019; Sevillya et al., 2020; Soucy et al., 2015).

The best way to identify an HGT event is by finding a phylogenetic incongruence which happens when there is a phylogenetic conflict between the species phylogeny and the gene phylogeny. If a gene was horizontally acquired, the branching pattern of the gene tree will not reproduce that of the species tree (Husnik & McCutcheon, 2017; Keeling & Palmer, 2008; Sevillya et al., 2020; Soucy et al., 2015) (Figure 1.1).

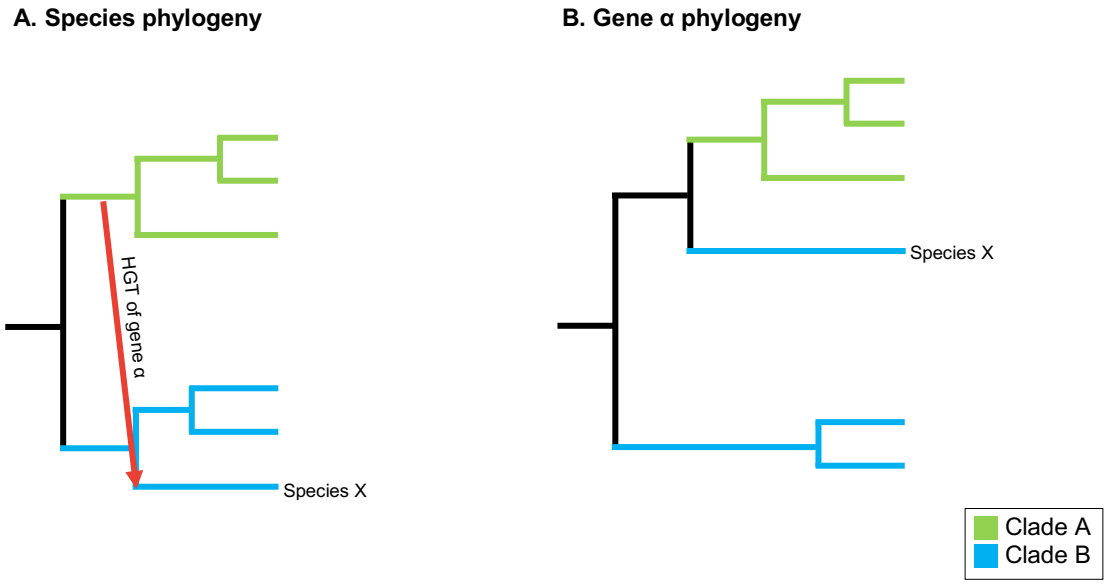


Figure 1.1. Representation of a hypothetical HGT event between two distinct clades. A) The species phylogeny places the different species in clade A (green) and clade B (blue). A hypothetical HGT event is represented by a red arrow and occurred when gene α from a species belonging to clade A was transferred to Species X, which belongs to clade B. B) The resulting gene α phylogeny, shows that as a consequence of the HGT, Species X clusters with the remaining species from clade A and not with species belonging to clade B. The branching pattern of the gene α does not reproduce that of the species tree, consisting on a phylogenetic incongruence, which suggests the occurrence of an HGT event.

The more closely related the donor and recipient species are, the more difficult it is to detect a phylogenetic incongruence and therefore, an HGT event. When detecting interdomain HGT events, the significant phylogenetical distance between donor and recipient species makes it easier to consubstantiate HGT. However, the similarity between the horizontally acquired gene and its donor organism tends to decrease over time (Sibbald et al., 2020). There are several other approaches to detect HGT events. These rely on differences between the composition of the genomes, namely the GC content, codon usage and the presence/absence and position of intron sequences in the analysed genes (Fitzpatrick, 2011; Sevillya et al., 2020; Yang & Rannala, 2016).

Currently, it is well established that foreign genes acquired by HGT are considered to be one of the most important sources of genome evolution among *Bacteria* and *Archaea*. The most common HGT mechanisms widely represented between prokaryotes are transformation, conjugation and transduction (Soucy et al., 2015) (Figure 1.2).

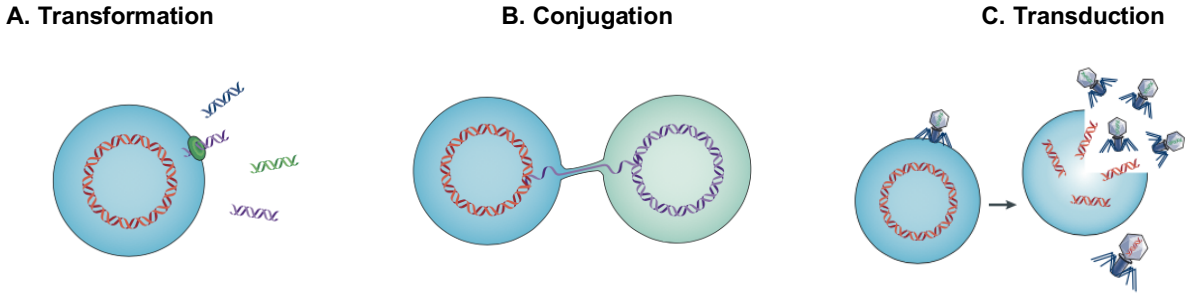


Figure 1.2. Mechanisms of HGT in Prokaryotes. A) Transformation consists in the uptake of foreign DNA from the environment by the host cell. B) Conjugation occurs when cell contact is established, and DNA is transferred from donor to recipient cell. C) Transduction is a virus-mediated form of DNA integration into the host cell (Adapted from Soucy et al., 2015).

1.1.1. Mechanisms of HGT in Eukaryotes

The significance and scale of HGT in *Eukarya* has been more controversial (Husnik & McCutcheon, 2017; Marcet-Houben & Gabaldón, 2010). Over the past decades, high-throughput sequencing technology along with comparative genomics tools, have become valuable resources for the identification of many genes of bacterial origin in eukaryotic genomes, that most likely arose from HGT. This new available data and increased knowledge shed a new light on the importance of HGT for eukaryotic evolution (Fitzpatrick, 2011; Husnik & McCutcheon, 2017).

Even though eukaryotes have several potential sources of exogenous DNA (Figure 1.3), the HGT process and its mechanisms remain unclear (Sibbald et al., 2020; Soucy et al., 2015). Different models have been proposed in order to explain this phenomenon. *Doolittle* hypothesized the '*ratchet model*' where it was explained that some eukaryotes possibly integrate the DNA from bacteria they phagocytize, into their genomes (Doolittle, 1998). The endosymbiotic theory is also emphasized by *Doolittle* since chloroplasts and mitochondria were originated from cyanobacteria and α-proteobacteria, respectively (Doolittle, 1998; Huang, 2013).

Huang later postulated the *'weak-link model'* that expands the former hypothesis for HGT and provides an explanation for other organisms that do not perform phagocytosis (Huang, 2013). This model proposes that foreign genes are more likely to be transferred to a certain organism at weakly protected stages of their life cycle, such as the unicellular stage. This is what makes unicellular organisms, as yeasts, more prone to horizontally acquire and transmit genes since the cell functions both as a somatic and germline cell (Husnik & McCutcheon, 2017; Marcet-Houben & Gabaldón, 2010).

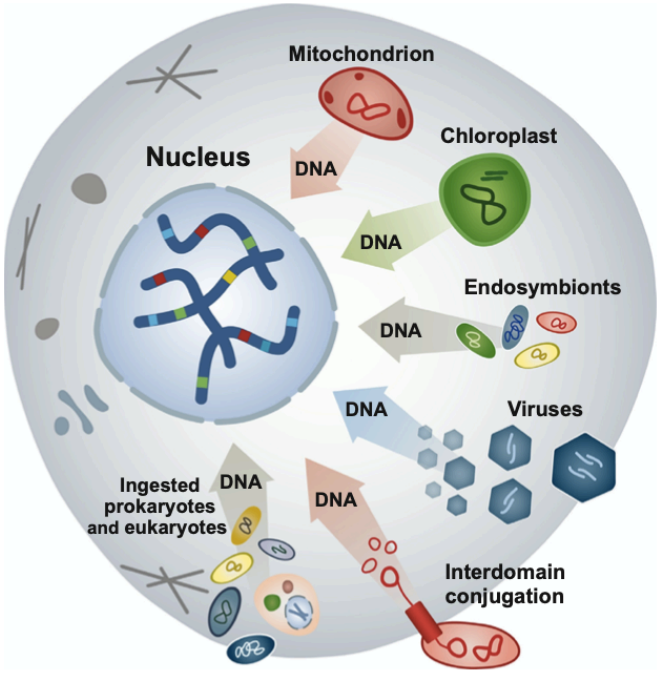


Figure 1.3. Possible sources of exogenous DNA contributing to HGT in Eukaryotes. The close and continued proximity of the eukaryotic nuclear DNA to endosymbionts and endosymbiont-derived organelles (mitochondria and chloroplasts) provides an excellent opportunity for the transferring and establishment of genes. This proximity is also achieved transiently when there is an ingestion of other prokaryotes and eukaryotes. It is also possible that the internalization of exogenous DNA can be mediated via bacteria (interdomain conjugation) and viruses (Sibbald et al., 2020).

The transfer of foreign DNA is more common when both donor and recipient organism cohabit in the same niche. However, in order for a gene to be fixed, it needs to go through a selection process. If the alien DNA does not confer any selective advantage to the recipient cell, it will most likely become a pseudogene or be completely lost. Interdomain transfer events also face more challenges since *Bacteria* and *Eukarya* have different gene structures and dissimilar transcription and translation mechanisms. Differences such as the spatial and temporal separation of transcription and translation in eukaryotes, or the production of polycistronic mRNA by prokaryotes must be overcome in order to establish a bacterial gene in a given eukaryotic organism (Gonçalves & Gonçalves, 2019; Husnik & McCutcheon, 2017; Kominek et al., 2019; Lindsey & Newton, 2019).

The fungal kingdom is the best sampled and most well-studied eukaryotic group. It represents a useful biological model regarding comparative genomics and evolution among eukaryotes (Fitzpatrick, 2011; Galagan et al., 2005; Husnik & McCutcheon, 2017). Some characteristics of fungal lifestyle may explain why these organisms are frequent recipient organisms, therefore owning a panoply of genes from

multiple sources. Most fungi are obligate osmotrophs that lead a symbiotic or saprophytic lifestyle, which means they are frequently associated or share their habitat with bacteria, respectively (Marcet-Houben & Gabaldón, 2010; Richards & Talbot, 2013). By being in continuous physical association with other organisms, the horizontal transfer of genes can be facilitated (Huang, 2003; Husnik & McCutcheon, 2017). Fungi are also able to break down complex biomolecules and feed on the resulting monomers, such as glucose and fructose. This confers adaptation to a wider range of environments and allows the spread of fungi (Richards & Talbot, 2013).

1.2. Alcoholic fermentation

Glucose and fructose are important carbon sources for cell metabolism. The metabolization of these carbohydrates starts with glycolysis, that is responsible for converting the monosaccharides into pyruvate, with concomitant NADH formation. Pyruvate can be further used as a substrate for respiration or fermentation. The alcoholic fermentation pathway involves the conversion of pyruvate to ethanol, along with CO_2 production, which allows the regeneration of $NAD(P)^+$ and the maintenance of the redox balance of the cell (Zamora, 2009) (Figure 1.4).

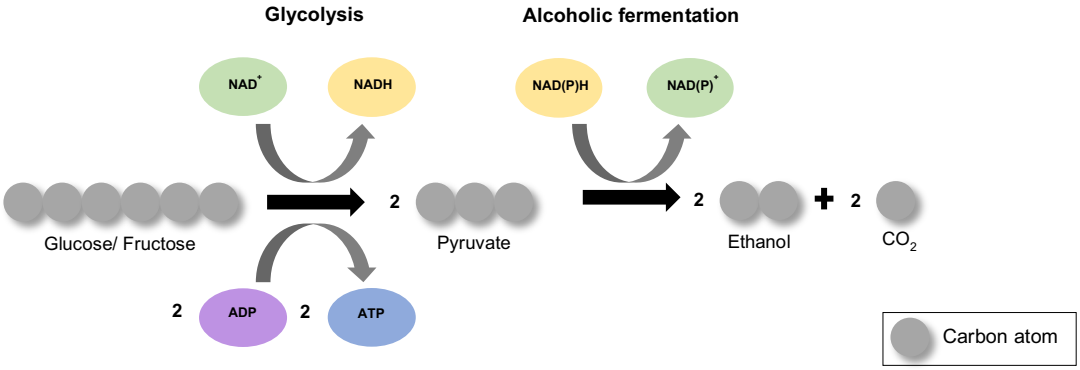


Figure 1.4. Glycolysis and the alcoholic fermentation pathway. The glycolytic pathway allows the conversion of one molecule of glucose or fructose into two molecules of pyruvate. This reaction forms two molecules of ATP and causes the reduction of NADH. Pyruvate is further used for alcoholic fermentation that regenerates the NAD^+ from glycolysis but also $NADP^+$, with concomitant CO_2 production.

Most yeasts are Crabtree-negative, which means that cellular respiration, which is an aerobic process, is the preferred metabolic pathway to obtain energy from carbohydrates (De Deken et al., 1966). In terms of energy yield, this pathway is more beneficial to the yeast than alcoholic fermentation. Per metabolized hexose, respiration produces a total of 36-38 molecules of ATP (Zamora, 2009). However, Crabtree-positive yeasts, such as *Saccharomyces cerevisiae*, preferentially engage in alcoholic fermentation, whether growing on aerobic or anaerobic conditions. In these yeasts, glucose is responsible for inhibiting cellular respiration. Therefore, this pathway is only activated when glucose levels are sufficiently low (De Deken et al., 1966; Zamora, 2009). Alcoholic fermentation not only allows the growth in anaerobic conditions, but it also confers adaptive advantage in certain niches, since ethanol can be potentially toxic and inhibit the growth of other species (Dashko et al., 2014).

1.2.1. The alcoholic fermentation pathway in *Saccharomyces cerevisiae*

Regarding the alcoholic fermentation pathway in *S. cerevisiae*, the pyruvate decarboxylase (Pdc1) enzyme is mostly responsible for catalysing the conversion of the end product of glycolysis, pyruvate, to acetaldehyde (Hohmann & Cederberg, 1990). Then, reduction of acetaldehyde is predominantly conducted by the alcohol dehydrogenase (Adh1), with concomitant NAD^+ regeneration (de Smidt et al., 2008) (Figure 1.5).

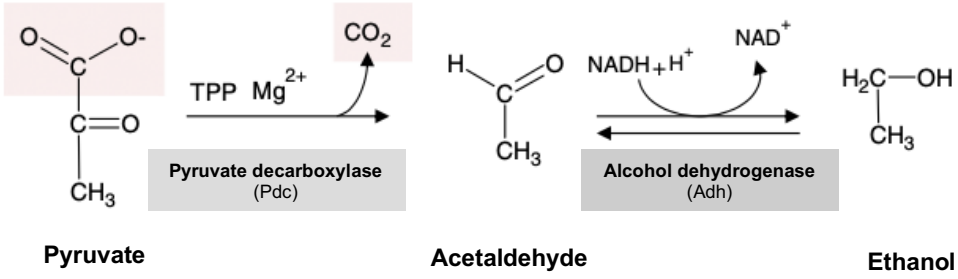


Figure 1.5. The alcoholic fermentation pathway in *Saccharomyces cerevisiae*. Pyruvate is converted into acetaldehyde with concomitant CO_2 production, by pyruvate decarboxylase (Pdc). Three Pdc enzymes are involved in this reaction (Pdc1, Pdc5 and Pdc6), but the catalysis is mainly performed by Pdc1. The cofactors of this reaction are magnesium (Mg^{2+}) and thiamine pyrophosphate (TPP). The last step is performed by the large family of alcohol dehydrogenases (Adh). In *S. cerevisiae*, this reaction is reversible. Adh1 is mainly involved in ethanol production, with concomitant NAD^+ formation. Adh2 catalyses the inverse reaction, consuming ethanol with concomitant formation of NADH (Adapted from Zea et al., 2015).

1.2.1.1. The pyruvate decarboxylase (Pdc)

In *S. cerevisiae* there are three structural genes that encode for active pyruvate decarboxylases (*PDC1*, *PDC5* and *PDC6*) (Hohmann & Cederberg, 1990; Hohmann, 1991). These are paralogue genes as the result of a whole genome duplication (WGD) event that occurred in an ancestor of *Saccharomyces*-related species (Gordon et al., 2009). It was observed that the transcription of both *PDC1* and *PDC5* is induced by the presence of glucose. The deletion of *PDC5* did not decrease the pyruvate decarboxylase specific activity of the cell. When *PDC1* was absent, *PDC5* was responsible for retaining 80% of the wild-type specific activity. The simultaneous deletion of both of these genes caused the total loss of ability to ferment glucose (Hohmann & Cederberg, 1990).

The *PDC6* gene was described later, and its deletion had no impact on the specific pyruvate decarboxylase activity of the wild type. It was observed that the *PDC6* gene was only transcribed in the presence of *PDC1* and preferentially non-fermentable substrates, such as ethanol. The role of this third *PDC* gene on the cell remains unclear (Hohmann, 1991). Only the mutants that did not produce the Pdc1 enzyme had a significant decrease of the enzyme specific activity (Hohmann and Cederberg, 1990; Hohmann, 1991), which means that in *S. cerevisiae*, Pdc1 is the main enzyme involved in the conversion of pyruvate to acetaldehyde.

1.2.1.2. The large family of alcohol dehydrogenases (Adh)

Alcohol dehydrogenases (Adh) are responsible for the last step of alcoholic fermentation, catalysing the interconversion of acetaldehyde and ethanol. Five *ADH* genes were originally described as the classical *ADH* in *S. cerevisiae* (*ADH1*, *ADH2*, *ADH3*, *ADH4* and *ADH5*). After the complete sequencing of the *S. cerevisiae* genome (Goffeau et al., 1996), more *ADH* genes were uncovered, such as the paralogues *ADH6* and *ADH7* and the *SFA1* gene. The respective enzymes differ in cell localization, expression profile, preferred cofactor and direction of reaction (de Smidt et al., 2008).

In *S. cerevisiae*, when fermentable substrates are available in the cell, Adh1 is the enzyme that preferentially catalyses the direct reaction, producing ethanol and restoring glycolytic NAD^+. When glucose concentration lowers and is no longer able to inhibit the respiratory pathway, the produced ethanol acts as a carbon source for aerobic respiration, and is therefore converted into acetaldehyde, preferentially by Adh2, with concomitant regeneration of NADH. Both Adh1 and Adh2 are NAD(H)-dependent enzymes (de Smidt et al., 2008, 2011).

The remaining *ADH* genes do not have a central role in alcoholic fermentation. However, the *ADH6* and *ADH7* paralogue genes encode for NADP(H)-dependent enzymes that are able to catalyse the inverse reaction, producing acetaldehyde, albeit with low efficiency. These two enzymes are described as cinnamyl dehydrogenases and are mostly involved in the synthesis of fusel alcohols and in ligninolysis (Larroy et al., 2002a, 2002b). The *SFA1* gene encodes a bifunctional NAD(H)-dependent alcohol and formaldehyde dehydrogenase which is mostly involved in the conversion of acetaldehyde to ethanol (Dickinson et al., 2003, Ida et al., 2012).

1.3. Fructophily

In *S. cerevisiae* and the majority of yeast species, glucose is the preferred carbon source for cell metabolism. However, in a small group of species, comprised by the distantly related *Zygosaccharomyces* genus and the *Wickerhamiella/Starmerella* clade (W/S clade), fructose is the preferred sugar. These species are called fructophilic. Fructophily is a rare metabolic trait among yeasts, and it consists on the preference for fructose over glucose (or any other hexose) as a carbon source, when both sugars are present (Cabral et al., 2015; Gonçalves et al., 2015, 2018; Leandro et al., 2014).

There seems to be a link between the fructophilic behaviour and the presence of a fructose-specific facilitator Ffz1 (Cabral et al., 2015; Leandro et al., 2014), that was originally acquired through HGT from a filamentous fungus by the MRCA of the W/S clade and later transferred from a *Wickerhamiella*-related species into the MRCA of the *Zygosaccharomyces* genus (Gonçalves et al., 2015). Fructophily appears to be a hallmark of species that inhabit the floral niche, such as the W/S-clade yeasts (Gonçalves et al., 2018), but also of a specialized group of lactic acid bacteria (Fructophilic lactic acid bacteria or FLAB) (Endo & Okada, 2008; Maeno et al., 2016, 2019).

1.3.1. The floral niche

The species from the non-conventional W/S clade inhabit fructose-rich environments, maintaining a strong ecological association with the floral niche. These microorganisms are mostly isolated from flowers and flower-visiting insects. On one hand, flowers provide the ideal microenvironment for the growth of ascomycetous yeasts, due to the sugar-rich nature of the nectar solution (de Vega et al., 2017; Lachance et al., 2001; Lachance, 2006). On the other hand, colonizing insects represent the main vectors of these fructophilic yeasts. While the majority of the members belonging to the *Wickerhamiella* genus are recovered from beetles and flies, *Starmerella* species are mostly carried by bees (Lachance et al., 2001). Each type of these colonizing insects harbours a very rich and specific microbiota, influenced by the unique combination of microorganisms that inhabit flowers where they often stopover (de Vega et al., 2017).

Besides W/S-clade species, Fructophilic Lactic Acid Bacteria (FLAB) are also strongly associated with this environment, being isolated from flower-visiting insects, such as honeybees, but also from beehives, fresh flowers and fruit peels (Endo & Salminen, 2013; Endo et al., 2009; Maeno et al., 2016). In addition to being fructophilic, both FLAB and W/S-clade species have another common trait, which is the ancestral loss of their fermentative capacities (Endo & Okada, 2008; Gonçalves et al., 2018; Maeno et al., 2016, 2019).

1.4. Loss of alcoholic fermentation in Fructophilic Lactic Acid Bacteria (FLAB)

The *Fructobacillus* and *Lactobacillus* species are the representatives of the FLAB group (Endo et al., 2018). These microorganisms do not produce ethanol since during the course of evolution, they have lost their ability to engage in alcoholic fermentation due to the lack of the bifunctional acetaldehyde-alcohol dehydrogenase gene (*adhE*) (Endo & Okada, 2008; Maeno et al., 2016, 2019). This gene encodes for an enzyme with alcohol dehydrogenase activity, converting acetaldehyde to ethanol with concomitant regeneration of $NAD(P)^+$. The absence of *adhE* in *Fructobacillus* spp. results in the accumulation of acetaldehyde and a shortage of $NAD(P)^+$ when glucose is used as a sole carbon source, so additional electron acceptors (such as fructose, oxygen or pyruvate) are required for an enhanced metabolism (Endo et al., 2013; Maeno et al., 2016, 2019).

A link between the loss of alcoholic fermentation and fructophily was previously demonstrated in *Fructobacillus fructosus*. The introduction of an *adhE* gene from the non-fructophilic bacterium *Leuconostoc mesenteroides*, not only allowed the re-establishment of alcoholic fermentation but it also obliterated the fructophilic behaviour (Maeno et al., 2019). This means that the loss of the *adhE* gene and consequently, alcoholic fermentation, may have occurred as the result of regressive evolution, in the process of the adaptation to fructose-rich environments (Endo et al., 2013; Maeno et al., 2019).

The reason why fructophily might be advantageous (besides the high availability of fructose in the floral niche), relies on the fact that this hexose acts both as a carbon source and as an electron acceptor (in

a reaction that converts fructose into mannitol). The metabolization of this hexose allows fructophilic bacteria to maintain redox balance by overcoming the shortage of NAD(P)$^+$. This is a key role for cell homeostasis that was mostly taken by alcoholic fermentation (Endo & Okada, 2008; Endo et al., 2009, 2012, 2013; Maeno et al., 2016, 2019) (Figure 1.6).

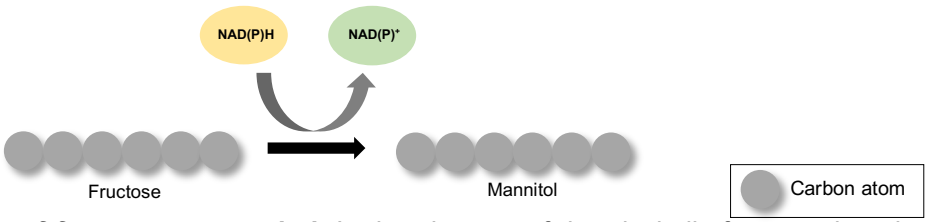


Figure 1.6. Conversion of fructose to mannitol. In the absence of the alcoholic fermentation, the glycolytic NAD(P)$^+$ is not regenerated, creating a shortage of this cofactor. Fructose, that functions both as a carbon source for cell metabolism and an electron acceptor, accepts the electrons from NAD(P)H and forms mannitol. This reaction allows the regeneration of NAD(P)$^+$ and the maintenance of the redox balance.

1.5. Loss of alcoholic fermentation in the ancestor of the *Wickerhamiella*/*Starmerella* (W/S) clade

The emergence of fructophily in W/S-clade yeasts seems also to be linked to the loss of alcoholic fermentation. Similarly to FLAB, the preference for fructose in these yeasts could be explained by the fact that the absence of alcoholic fermentation (that translates in the loss of both pyruvate decarboxylase and alcohol dehydrogenase enzymes) likely imposed a redox imbalance (Gonçalves et al., 2018).

Contrary to glucose, fructose can be directly converted into mannitol by a mannitol dehydrogenase enzyme (Mtdh). This reaction restores NADP$^+$ and is therefore hypothesized that it could have helped to maintain redox balance in the W/S ancestor where alcoholic fermentation was lost (Gonçalves et al., 2019). Even though the deletion of *MtDH1* and its paralogue *MtDH2* in *Starmerella bombicola* does not obliterate the fructophilic behaviour, it significantly improves glucose consumption while decreasing fructose consumption, therefore attenuating the fructophilic profile. This means that a significant fraction of fructose is being converted to mannitol (Gonçalves et al., 2019).

The '*less-is-more*' hypothesis refers that gene losses are the result of an evolutionary adaptive response to the occupation of new environments. They occur frequently among different species and are strong drivers of phenotypical diversity (Albalat & Cañestro, 2016). For example, in some subpopulations of *S. cerevisiae*, the ability to produce aquaporins was lost, due to the accumulation of nonsense mutations, as they adapted to high-sugar environments (Will et al., 2010).

1.5.1. Reacquisition of alcoholic fermentation in the W/S clade

Regressive evolution is a common evolutionary process in the subphylum Saccharomycotina (yeasts), that comprises the W/S clade. It was inferred that the MRCA of this subphylum was metabolically more complex than its descendants, which showed a general tendency for the extensive loss of metabolic traits (Shen et al., 2018). However, it was also observed that even though HGT events rarely occurred

in the Saccharomycotina, the W/S clade is proving to be a group with a higher incidence of these events. The horizontal acquisition of genes in the W/S clade has allowed the re-establishment of once lost features (Gonçalves & Gonçalves, 2019; Gonçalves et al., 2018, 2020; Shen et al., 2018).

The loss followed by reacquisition of metabolic pathways is also observed within other fungi outside the subphylum Saccharomycotina. Two microsporidian parasites horizontally acquired genes from distantly related prokaryotic and eukaryotic donors, in order to reconstruct incomplete metabolic pathways. The authors referred that this mode of acquisition raised important questions, since it was not clear why these were lost in the first place (Pombert et al., 2012).

There are several examples of metabolically important genes acquired by HGT in the W/S clade, such as the bacterial extracellular invertase *SacC*. The SacC protein is responsible for hydrolysing sucrose to fructose and glucose (Gonçalves et al., 2018). Some species of the W/S clade were even able to acquire entire bacterial operons in order to be able to biosynthesize the essential vitamin B1 (Gonçalves & Gonçalves, 2019) and siderophores (Kominek et al., 2019). These represent remarkable examples of interdomain HGT events, since operons are complex prokaryotic structures that these yeasts were able to adapt, for them to become functional in the eukaryotic setting (Gonçalves & Gonçalves, 2019; Kominek et al., 2019).

HGT events were also involved in the reacquisition of the alcoholic fermentation pathway, that occurred in some of the W/S-clade species, contrary to FLAB that never restored this pathway. The re-establishment of the pathway occurred through the recruitment of a new decarboxylase (Aro10), that took over the role of Pdc1 and by the capturing of bacterial *ADH* genes (Gonçalves et al., 2018, 2020).

Curiously, similarly to FLAB, these yeasts continued to produce mannitol and even developed a new strategy to do so when glucose is the sole carbon source available (Gonçalves et al., 2019). While in fructophilic bacteria the production of mannitol from fructose seems to be linked to redox balance, it is possible that in W/S-clade yeasts, once the alcoholic fermentation was re-established, mannitol has evolved to fulfil different roles on the cell, namely as a thermal protector (Gonçalves et al., 2019).

1.5.1.1. Reinstatement of pyruvate decarboxylase (Pdc1) activity

The evidence for the loss of the *PDC1* gene from W/S-clade genomes was obtained while searching for orthologues using the S. cerevisiae *PDC1* sequence as query (Gonçalves et al., 2018). Except for one species (*Wickerhamiella versatilis*), *PDC* genes seem to be absent in W/S-clade yeasts. Moreover, it was shown that the *PDC*-like genes found in *W. versatilis* genome came from bacteria. The closest gene found in other W/S-clade species is a phenylpyruvate decarboxylase (*ARO10*) (Gonçalves et al., 2018, 2020) (Figure 1.7).

In *S. cerevisiae* Aro10 presents a broad substrate specificity, despite having a very low affinity to pyruvate as a substrate (Kneen et al., 2011), and being therefore not involved in alcoholic fermentation.

However, it was observed that the deletion of *ARO10* in *St. bombicola* resulted in the loss of the ability to produce ethanol and therefore engage in alcoholic fermentation. These results indicate that the loss of *PDC1* in the W/S clade might have been counteracted by the acquisition of a new function by the related decarboxylase Aro10 (Gonçalves et al., 2018).

1.5.1.2. Reinstatement of alcohol dehydrogenase (Adh) activity

The loss of native *ADH* genes from W/S-clade genomes was supported after a Maximum Likelihood (ML) phylogeny clustered all of the *ADH*-like sequences from these species with bacteria and not fungi (Gonçalves et al., 2018). Besides this, in at least one genome (*Wickerhamiella galacta*), the *ADH1* gene is absent. According to the ML phylogeny of Adh1, three independent HGT events were involved in the reacquisition of *ADH1* in the W/S clade from three different groups of bacteria (Gonçalves et al., 2018, 2020) (Figure 1.7).

Concerning Adh6, the ML phylogeny also placed the W/S-clade sequences with bacteria, and it was possible to identify one single HGT event (Figure 1.7). It is also observed that *ADH6* xenologues have duplicated several times within each W/S-clade species, *Starmerella bacillaris* and *Starmerella magnoliae* being the ones with the most paralogous genes (Gonçalves et al., 2018, 2020).

Regarding the role of these Adh enzymes in alcoholic fermentation, it was observed that elimination of *ADH1* in *St. bombicola*, impairs both ethanol production and assimilation. However, elimination of each of the two *ADH6* paralogues did not produce any critical differences concerning ethanol metabolism. This means that the Adh1 from *St. bombicola* is the main enzyme involved on the alcoholic fermentation pathway and that, contrary to what was observed in *S. cerevisiae*, whose Adh1 mostly catalyses the direct reaction (de Smidt et al., 2008, 2011), in *St. bombicola* this enzyme is involved in both reactions (producing and assimilating ethanol). Adh6 is possibly fulfilling a substitution role when Adh1 is missing (Gonçalves et al., 2018). This functional interchangeability between Adh proteins is also observed in *S. cerevisiae* deletion mutants (de Smidt et al., 2011).

Concerning the preferred cofactors, for *St. bacillaris*, *St. bombicola* and *St. magnoliae* cell extracts, Adh activity was measured by either adding NADH or NADPH to the reaction mixture. The measured NAD(P)H enzymatic activities, at least in *St. bombicola*, are most likely from Adh1 and not Adh6. This is because when the *ADH1* from *St. bombicola* was eliminated, NADH and NADPH activities were no longer observed in cell-free extracts (Gonçalves et al., 2018). The NADPH-dependent activity in Adh1 is an uncommon characteristic in yeasts, since all of the non-W/S-clade yeasts evaluated so far, including *S. cerevisiae,* were only able to use NADH as a cofactor (de Smidt et al., 2008, 2011; Gonçalves et al., 2018). Therefore, contrary to other yeast Adh1 enzymes, the Adh1 from *St. bombicola* is able to perform not only the interconversion of acetaldehyde and ethanol, but also has the ability to use NADH and NADPH as cofactors (Gonçalves et al., 2018).

1.5.2. Phylogenetic relationships and HGT dynamics in the W/S clade

The distribution of species throughout the *Wickerhamiella* (de Vega et al., 2017) and *Starmerella* (Santos et al., 2018) genera was based on D1/D2 and ITS phylogenies. Even though the D1/D2 and ITS regions are important biomarkers for yeast taxonomic identification (Tian et al., 2015), these single gene phylogenies may not as precisely represent the complexity of the history of evolutionary events. For this reason, multigene or even phylogenomic approaches (Figure 1.7) are considered to be a better way to obtain a robust and accurate representation of evolutionary events across time, since it is possible to study an entire (or a significant portion of a) proteome (Feng et al., 2017; Robbertse et al., 2006).

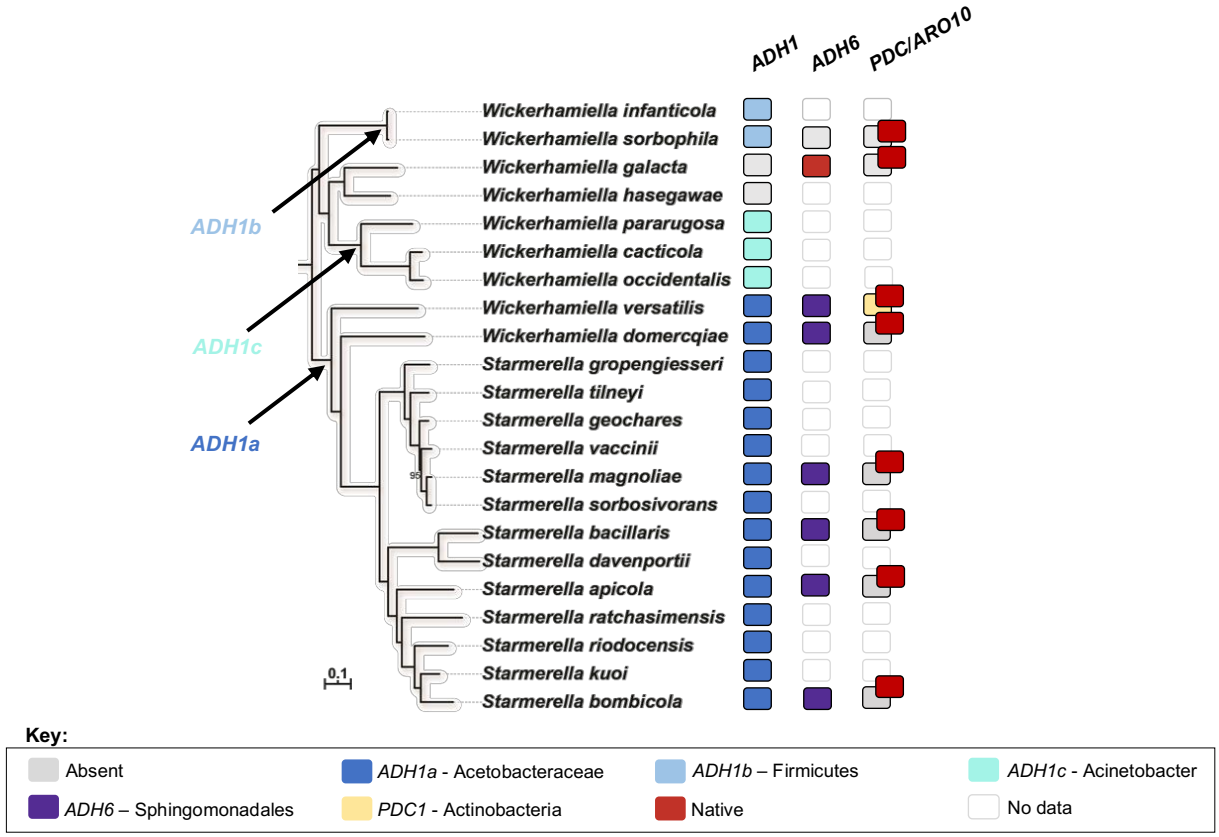


Figure 1.7. Phylogenetic relationships and HGT dynamics in the W/S clade. According to the phylogenomic analysis, *Wickerhamiella versatilis* and *Wickerhamiella domercqiae* are placed within the *Starmerella* subclade instead of the *Wickerhamiella* subclade. The possible branches where the acquisition of *ADH1a*, *ADH1b* and *ADH1c* of each subgroup occurred are marked by arrows. The gene donors of the horizontally transferred genes (*ADH1a*, *ADH1b*, *ADH1c*, *ADH6* and the *PDC1* from *W. versatilis*) are labelled by colours and their distribution is presented in the ML phylogeny. The presence of the native *ARO10* as well as the absence of the native *PDC1* in all the evaluated genome is also labelled in the ML phylogeny (Gonçalves et al., 2018, 2020). There are still several W/S-clade species yet to characterize and the information regarding their genes is absent (Adapted from Gonçalves et al., 2020).

As a result of the phylogenomic analyses (Figure 1.7), it has been revealed that *W. domercqiae* and *W. versatilis* are in fact, more closely related to the *Starmerella* genus than to the *Wickerhamiella* genus (Gonçalves et al., 2020, Shen et al., 2018), as the D1/D2 phylogenies initially described (de Vega et al.,

2017). The phylogenomic approaches are crucial to better establish phylogenetic relationships and further explore the HGT dynamics in the W/S clade.

1.5.2.1. The independent HGT events of bacterial *ADH1* and *ADH6*

The acquisition of *ADH1* in the W/S clade occurred independently, at least three times, during the course of evolution. This means that there are three different donor organisms and therefore three distinct *ADH1* gene types (*ADH1a*, *ADH1b*, *ADH1c*). According to the available information, it was inferred that an *ADH1* from an Acetobacteraceae-related species was horizontally transferred to the MRCA of the *Starmerella* and *Starmerella*-related species (*W. domercqiae* and *W. versatilis*). The *ADH1* from this lineage was named as *ADH1a* (Gonçalves et al., 2018, 2020) (Figure 1.7).

An *ADH1* from a Firmicutes-related species was inferred to be horizontally transferred to the MRCA of some *Wickerhamiella* species (as *W. infanticola* and *W. sorbophila*) and it was named *ADH1b*. The *ADH1* from an Acinetobacter-related species was inferred to be horizontally transferred to the MRCA of another W/S-clade subgroup (*W. cacticola*, *W. occidentalis* and *W. pararugosa*) and was named as *ADH1c* (Gonçalves et al., 2020). Concerning the *ADH6* gene, it was observed that some species from the *ADH1a* lineage have received an *ADH6* from a species belonging to the Sphingomonadales order. It was still not clear whether this (or other) bacterial *ADH6* genes were also present in the species of other subgroups, or if *ADH6* is absent from most genomes (Gonçalves et al., 2018, 2020) (Figure 1.7).

The Acetobacteraceae family is often associated to flowers and flower-visiting insects (Iino et al., 2012). Since the W/S-clade species are also associated to these environments (Lachance et al., 2001), the close and continued proximity between donor and recipient species have provided a great opportunity for these HGT event to occur (Huang, 2003; Husnik & McCutcheon, 2017).

1.5.2.2. The '*loss followed by reacquisition*' hypothesis

There are two hypotheses that can explain the order of events that took place in the MRCA of the W/S clade. These events are the loss of the alcoholic fermentation genes and the acquisition of this pathway. The '*acquisition preceded loss*' hypothesis postulates that the HGT events occurred before the loss of the native *ADH* genes. There are some useful clues to why this might not have occurred. Firstly, no W/S-clade genomes have been found with both the native and foreign *ADH* genes that could represent an intermediate stage of the loss. Secondly, there is a lineage without *ADH1* genes, comprised by *Wickerhamiella galacta* and *Wickerhamiella hasegawae* (Figure 1.7). It is more likely that this lineage (*ADH0*) arose by the absence of an HGT event on its MRCA, than the occurrence of an independent loss of *ADH1*. In line with this, besides not having any *ADH1* gene, *W. galacta*, and all the other evaluated W/S-clade species (with the exception of *W. versatilis*) do not have a *PDC1* gene (Gonçalves et al., 2018, 2020).

Altogether, these findings strongly support the '*loss followed by reacquisition*' hypothesis. This hypothesis proposes that the ancestral loss of the alcoholic fermentation pathway, occurred before the acquisition of bacterial *ADH1* genes. This is also highlighted by the fact that at least in *St. bombicola*, Aro10 has proposedly evolved to be able to engage in alcoholic fermentation, occupying the role of the once lost Pdc1 (Gonçalves et al., 2018, 2020). Nevertheless, it is still not clear why the W/S-clade yeasts horizontally acquired the foreign *ADH1* genes after losing the native ones. It could have happened that the need for the re-establishment of alcoholic fermentation re-emerged in W/S-clade yeasts.

1.6. Aims of the project

It is currently estimated that the W/S clade comprises about 90 acknowledged species, 43 belonging to *Wickerhamiella* (de Vega et al., 2017) and the remaining 47 to *Starmerella* (Santos et al., 2018). Moreover, about 50 putative new species are yet to be formally described and characterized (Gonçalves et al., 2020). Therefore, so far, the set of events uncovered are based on a still limited amount of genomic data.

For this project, the main goal was to build on the previous work regarding alcoholic fermentation in the W/S clade by uncovering more HGT events involving the *PDC1*, *ADH1* and *ADH6* genes. To achieve this, tBLASTx searches, using *S. cerevisiae PDC1*, *ADH1* and *ADH6* genes as queries, were performed against 26 W/S-clade genomes. The reconstruction of ML phylogenies that contained this new information was crucial to better elucidate the different patterns of horizontal acquisitions of these genes and to give more support to the phylogenomic analysis of the W/S clade.

Understanding the impact that these genes have in alcoholic fermentation was also a goal of this project. To accomplish that, a set of species from each subgroup, that better represent the biodiversity of this clade, were selected and ethanol production and assimilation were investigated. This data allowed a better understanding regarding the preferred direction of the interconversion of acetaldehyde and ethanol. These results were complemented with enzymatic assays on protein extracts to provide a preliminary picture of the alcohol dehydrogenase activities among W/S-clade species. The purification of Adh1 enzymes from *St. bombicola* (Adh1a) and *W. cacticola* (Adh1c) was also envisioned, in order to thoroughly characterize substrate and cofactor specificities.

After gathering all the results, it was possible to correlate the presence of the alcoholic fermentation genes, the ability to produce and/or assimilate ethanol and the cofactor preference within each subgroup. This further allowed a broader evolutionary perspective of alcoholic fermentation evolution in the W/S clade.

2. Materials and Methods

2.1. Strains and culture conditions

The W/S-clade species that were used in the experiments were the following: *Starmerella apicola* PYCC 3042; *Starmerella bacillaris* PYCC 3044; *Starmerella davenportii* CBS 9069; *Starmerella floris* PYCC 8435; *Starmerella geochares* PYCC 8323; *Starmerella gropengiesseri* PYCC 2915; *Starmerella kuoi* NRRL Y-27208; *Starmerella lactis-condensi* PYCC 8434; *Starmerella ratchasimensis* PYCC 7052; *Starmerella riodocensis* PYCC 8433; *Starmerella sirachaensis* PYCC 7054; *Starmerella sorbosivorans* PYCC 8429; *Starmerella tilneyi* CBS 8794; *Starmerella vaccinii* PYCC 8432; *Wickerhamiella alocasiicola* PYCC 8427; *Wickerhamiella bombiphila* PYCC 8430; *Wickerhamiella cacticola* PYCC 6392; *Wickerhamiella domercqiae* PYCC 3067; *Wickerhamiella galacta* PYCC 8318; *Wickerhamiella hasegawae* PYCC 8324; *Wickerhamiella infanticola* PYCC 8312; *Wickerhamiella jalapaonensis* PYCC 8424; *Wickerhamiella kazuoi* PYCC 8329; *Wickerhamiella kurtzmanii* PYCC 8437; *Wickerhamiella nectarea* PYCC 8436; *Wickerhamiella occidentalis* PYCC 6399; *Wickerhamiella pararugosa* PYCC 6791; *Wickerhamiella parazyma* PYCC 8426; *Wickerhamiella siamensis* PYCC 7069; *Wickerhamiella slavikovae* PYCC 8320; *Wickerhamiella spandovensis* PYCC 8431; *Wickerhamiella vanderwaltii* PYCC 3671.

The W/S-clade species (*St. apicola*; *St. bacillaris*; *W. cacticola*; *W. domercqiae*; *W. galacta*; *W. hasegawae*; *W. infanticola*; *W. jalapaonensis*; *W. kazuoi*; *W. nectarea*; *W. occidentalis*; *W. pararugosa*; *W. parazyma*; *W. siamensis*; *W. vanderwaltii*), *Candida incommunis* PYCC 4837 and *Acetobacter malorum* PYCC 8266 were obtained from PYCC (Portuguese Yeast Culture Collection, Caparica, Portugal). While W/S-clade species and *C. incommunis* were maintained in Yeast Malt Agar (YMA medium) [0,5% (w/v) Peptone (BD Biosciences); 0,3% (w/v) Yeast Extract (BD Biosciences); 0,3% (w/v) Malt Extract (BD Biosciences); 1% (w/v) D-Glucose (LABChem); 2% (w/v) Agar (LABChem)], *A. malorum* was cultivated in GYPA (Glucose-Yeast-Peptone Agar) [5% (w/v) D-Glucose; 0,5% (w/v) Yeast Extract; 0,3% (w/v) Peptone].

2.2. Reconstruction of Maximum Likelihood (ML) Phylogenies of Adh1, Adh6 and Pdc1/Aro10 proteins

2.2.1. Genome sequencing of W/S-clade species

The 26 W/S-clade sequenced genomes that were used in this project were previously obtained from two different sources. The 14 following genomes were sequenced in the context of the Y1000+ Project that is focused on the whole genome sequencing of all known species of the Subphylum Saccharomycotina; http://y1000plus.org (Hittinger et al., 2015): *St. davenportii*; *St. geochares*; *St.*

gropengiesseri; *St. kuoi*; *St. ratchasimensis*; *St. riodocensis*; *St. sorbosivorans*; *St. tilneyi*; *St. vaccinii*; *W. cacticola*; *W. hasegawae*; *W. infanticola*; *W. occidentalis*; *W. pararugosa*. The 12 subsequent genomes were sequenced in the laboratory: *St.* floris; St. *lactis-condensi*; *St. sirachaensis*; *W. alocasiicola*; *W. bombiphila*; *W. jalapaonensis*; *W. kurtzmanii*; *W. nectarea*; *W. parazyma*; *W. siamensis*; *W. slavikovae*; *W. spandovensis*.

New genomes were previously obtained in the host laboratory. Briefly, DNA from overnight grown cultures and paired-end Illumina MiSeq 250 bp genomic reads were further obtained after 500 sequencing cycles. The raw reads were first pre-processed by trimming of adapters and low-quality bases using Trimmomatic v0.33 (Bolger et al., 2014). The optimal k-mer length for each genome's assembly was calculated using KmerGenie v1.6982 (Chikhi & Medvedev, 2014). The processed reads were used to generate *de novo* assemblies using SPADES v3.7.0 (Bankevich et al., 2012).

2.2.2. Phylogenomic analysis of the W/S clade

A phylogenomic tree was previously constructed using a total of 380 orthogroups present in single copy in all genomes analysed (Gonçalves et al., 2021, unpublished). Briefly, single copy orthologues present in, at least 90% of the species, were retrieved using Orthofinder 2 (Emms & Kelly, 2019) and a concatenated alignment was obtained in MAFFT v7.453 (Multiple alignment program for amino acid or nucleotide sequences) (Katoh & Standley, 2013) using the L-INS-i strategy ('--localpair'). The L-INS-i method allows the alignment of sequences with common regions while surrounded by non-alignable domains. This alignment was used to infer a Maximum Likelihood (ML) tree using IQTREE v2.0 (Nguyen et al., 2014) with an automatic detection for the best-fitting model of amino acid evolution and 1.000 ultrafast bootstrapping replicates (Minh et al., 2013).

2.2.3. Identification of *ADH1*, *ADH6* and *PDC1/ARO10* genes in W/S-clade species

The *Saccharomyces cerevisiae ADH1* (Y0L086C), *ADH6* (YMR318C) and *PDC1* (YLR044C) gene sequences were retrieved from the *Saccharomyces* genome database (SGD). A custom query consisting of *ADH1*, *ADH6* and *PDC1* genes from *S. cerevisiae* was used in tBLASTx searches against the 26 genomes of W/S-clade species. The best tBLASTx hits (*e*-value cutoff < $1e^{-5}$) were subsequently analysed. The establishment of a low E-value (Expect value) cutoff is crucial to avoid the random background noise caused by less significant hits (Mitrophanov & Borodovsky, 2006).

2.2.4. Protein prediction of Adh1, Adh6 and Pdc1/Aro10

The recovered DNA sequences (2 kb upstream and downstream of the candidate gene location) were further analysed in AUGUSTUS (Stanke et al., 2008) in order to predict the protein sequences, using *S. cerevisiae* as a model organism. The protein sequences were blasted against the NCBI non-redundant protein database (BLASTp) in order to confirm orthology with the proteins of interest (Adh1, Adh6 or

Aro10/Pdc1). Whenever the best hits corresponded to an alcohol dehydrogenase or a pyruvate decarboxylase, the respective protein and coding sequences (CDS) were retrieved.

2.2.5. Construction of preliminary Adh1 and Adh6 Maximum Likelihood (ML) phylogenies

Since Adh1 and Adh6 proteins belong to the same family of enzymes and are therefore similar (de Smidt et al., 2008), the construction of preliminary ML phylogenies was performed in order to distinguish between the two groups of sequences. To construct the preliminary Adh1 and Adh6 trees, the top 5.000 BLASTp hits were retrieved from NCBI (non-redundant database, as of July 2020). These hits were recovered by using *Starmerella bombicola* proteins as queries (Adh1 in the case of the preliminary Adh1 tree; one of the two Adh6, in the case of the preliminary Adh6 tree).

To eliminate redundant sequences, CD-HIT v4.8.1 (Cluster database at High Identity with Tolerance) (Li & Godzik, 2006) was subsequently used to cluster sequences according to a 95% similarity threshold. Proteins with 95% sequence identity were removed, and each cluster ended up with one representative sequence. Following this, protein sequences were aligned with MAFFT v7.453 using the L-INS-i method ('--localpair'). Poor alignment segments were trimmed using trimAl v1.2 (Capella-Gutiérrez et al., 2009), a tool for automated alignment trimming. The 'gappyout' method was selected in order to remove gap-rich regions of the aligned sequences. The preliminary phylogenetic trees were constructed using IQ-TREE v2.0, using an ultrafast bootstrap method (-bb) and analysed via iTOL v5.0 (interactive tree of life, https://itol.embl.de) (Letunic & Bork, 2021), where the mid-point root and bootstrap values (>90%) were accessed. The sequences that formed an outgroup in the Adh1 and Adh6 preliminary phylogenies were removed as indicated in Figures A1 and A2 from the Appendix, respectively. The remaining sequences were kept for downstream analysis.

2.2.6. Construction of Maximum Likelihood (ML) phylogenies of Adh1, Adh6 and Pdc1/Aro10

The selected Adh1 and Adh6 sequences were added to the respective datasets from Gonçalves et al., 2018, to reconstruct the final ML phylogenies. These were subsequently clustered with CD-HIT v4.8.1 (95% sequence identity for Adh1 and 98% identity for Adh6). The sequences were aligned and trimmed as aforementioned. As for the Pdc retrieved sequences, they were also added to the respective dataset from Gonçalves et al., 2018. The alignment and the trimming were subsequently performed, as mentioned before. The final phylogenetic trees were reconstructed with IQ-TREE v2.0 (67) using the LG+I+G4 model of substitution (found as the best fitting model for all of the four alignments) and ultrafast bootstrap (-bb 1.000) (68) for branch support determination. Phylogenies were mid-point rooted and visualized in iTOL v5.0.

2.3. Ethanol production and assimilation in W/S-clade species

2.3.1. Growth conditions

To study ethanol production and assimilation profiles in the W/S clade, a subset of *Wickerhamiella* and *Starmerella* species were selected to represent the diversity of the clade: *St. apicola*; *St. bacillaris*; *W. cacticola*; *W. domercqiae*; *W. galacta*; *W. hasegawae*; *W. infanticola*; *W. jalapaonensis*; *W. kazuoi*; *W. nectarea*; *W. occidentalis*; *W. pararugosa*; *W. parazyma*; *W. siamensis*; *W. vanderwaltii*.

2.3.1.1. Ethanol assimilation tests

The assimilation of ethanol was determined by the macroscopic observation of growth in minimal medium YNB (Yeast Nitrogen Base) (DIFCO), supplemented with 2% (v/v) Ethanol (Honeywell). The YNB medium is used for nutrient assimilation studies since it lacks a carbon source that is further added. The YNB medium, supplemented with 2% (w/v) D-Glucose was used as a positive control in this experiment.

Each W/S-clade species was pre-cultivated in 50 mL Erlenmeyer flasks, containing 10 mL of YNB + 0,5% (w/v) D-Glucose. This medium has a lower glucose concentration to avoid hexose storage in cells, which could influence the further growth observations in the YNB + 2% (v/v) Ethanol. The cells were incubated at 25°C overnight with orbital shaking. On the following day, cell cultures were observed by optical microscopy in order to check for possible contamination. Cell densities were measured in the spectrophotometer (Biochrom) and the pre-inoculum was diluted to an OD_{640}=0,05, in two 250 mL Erlenmeyers, each containing 30 mL of YNB medium, one supplemented with 2% (w/v) D-Glucose and the other with 2% (v/v) Ethanol. Cultures were cultivated at 25°C for up to 25 days with orbital shaking. When it took more than 15 days for the growth to be observable, the growth was considered ‘delayed’. When the culture did not grow vigorously (low turbidity) in this medium, the growth was considered ‘weak’.

2.3.1.2. Measurement of ethanol production by HPLC

The measurement of extracellular concentrations of ethanol (g L^{-1}) was performed by high performance liquid chromatography (HPLC) in cultures that grew on 2x YP (Yeast-Extract Peptone) [2% (w/v) yeast extract; 4% (w/v) peptone], supplemented with 10% (w/v) D-Fructose (VWR) and 10% (w/v) D-Glucose (20FG medium). This medium has a high sugar concentration, which mimics the floral niche and is based on rich medium YP (and not on minimal YNB) in order to avoid nutrient exhaustion that could hamper the sugar uptake (Cabral et al., 2015).

For the pre-growth, each culture was pre-cultivated in 50 mL Erlenmeyer flasks, containing 10 mL of 20FG medium. These cultures were incubated at 25°C with orbital shaking and the growth was monitored for ~400 h, by determining the OD_{640nm}. During that interval microscopical observation of

cultures was performed to guarantee the absence of contamination and 800 μL aliquots of culture were collected at different time-points for HPLC analysis. The 800 μL aliquots were transferred to 1,5 mL tubes and immediately centrifuged for 5 min at 13.000 x g. The supernatant was recovered and filtered using a 2 mL syringe (BBraun) with a 13 mm Nylon filter (Filter-Lab). The samples were stored at -20°C until being used for HPLC analyses. An aliquot of 20FG medium was also retrieved for HPLC analyses as it represents T0.

Concerning *St. apicola*, *St. bacillaris*, *W. domercqiae*, *W. galacta*, *W. hasegawae*, *W. infanticola*, *W. kazuoi*, *W. nectarea* and *W. vanderwaltii*, the quantification was performed by using a carbohydrate analysis column (300 mm x 7,8 mm, Aminex HPX-87P; Biorad) and a differential refractometer (LKB 2.142). The column was kept at 80°C and H_2O was used as the mobile phase at 0,6 ml min^{-1}. Regarding *W. cacticola*, *W. jalapaonensis*, *W. occidentalis*, *W. pararugosa*, *W. parazyma* and *W. siamensis*, the quantification was performed by utilizing a Thermo CarboPac column (250 mm x 4,0 mm, Dionex ICS3000; ThermoFisher-Dionex) with pulsed amperometric detection. The column was kept at 25°C and 612 mM NaOH was used as the mobile phase at 0,4 ml min^{-1}. Ethanol production was considered when the measurable amounts were superior to 5,00 g/L.

2.4. Enzymatic assays

2.4.1. Growth conditions

To better characterize Adh activity in W/S-clade species (namely cofactor preference), the following species were selected in order to adequately represent the diversity of the clade: *St. apicola*; *St. bacillaris*; *W. domercqiae*; *W. galacta*; *W. hasegawae*; *W. jalapaonensis*; *W. kazuoi*; *W. vanderwaltii*. Besides W/S-clade species, one non-W/S-clade yeast, *C. incommunis* and one bacterium, *A. malorum* were also selected for the experiment. These species were pre-cultivated in 50 mL Erlenmeyers, each containing 10 mL of medium. All yeast species were cultivated in rich medium YPD (Yeast-Extract-Peptone-Dextrose) [1% (w/v) Yeast Extract; 2% (w/v) Peptone; 2% (w/v) Glucose]. The species *A. malorum* was cultivated in GYP (Glucose-Yeast-Peptone) [5% (w/v) glucose; 0,5% (w/v) Yeast Extract; 0,3% (w/v) Peptone]. Cultures were pre-grown at 25°C overnight with orbital shaking. On the following day, optical densities were measured and pre-inoculums were diluted to an OD_{640}=0,1 to prepare the inoculum in 250 mL Erlenmeyers containing 50 mL of the respective medium.

2.4.2. Preparation of cell-free extracts for enzymatic assays

After the cultures reached mid-exponential phase (after ~24 h of growth), cells were harvested by centrifugation. Cultures were distributed to 50 mL falcons and centrifuged for 10 min, 8.500 x g at 4°C. The supernatant was discarded and pellets were washed twice with 30 mL of TRIS buffer [50 mM triethanolamine hydrochloride (pH=7,6) (NZYTech); 1 μM of Phenylmethanesulfonyl fluoride (PMSF) (Thermo Scientific)] following the same centrifugation conditions. The pellets were always kept on ice.

Pellets were resuspended in 1 mL of TRIS buffer and 500 μL were distributed to two 2 mL tubes, containing glass beads. The mixture was centrifuged for 1 min, at 16.200 x g at room temperature (RT) and the supernatant was discarded. Following this, 400 μL of TRIS Lysis Buffer [0,1M triethanolamine hydrochloride; 1 μM PMSF; 1 mM Dithiothreitol (DTT) (Thermo Scientific)] was added to the tubes and mixed in order to resuspend cells. The cells were disrupted by six alternating cycles of 1 min of vortex and 1 min of cooling on ice. Cell debris were removed by centrifugation for 20 min, 16.200 x g at 4°C.

2.4.3. Alcohol dehydrogenase enzymatic assay

The freshly prepared crude extracts were used for the enzymatic assays. In this experiment, the reversible reaction catalysed by the large family of alcohol dehydrogenase (Adh) enzymes was evaluated. The direct reaction involves the reduction of acetaldehyde to ethanol, along with the oxidation of NAD(P)H to $NAD(P)^+$. The inverse reaction involves the oxidation of ethanol to acetaldehyde, with the concomitant reduction of $NAD(P)^+$ to NAD(P)H.

To measure the direct reaction, 50 mM of phosphate buffer [50 mM K_2PO_4, 50 mM KH_2PO_4 (pH=7,6)], 1 mM of NADH (Sigma-Aldrich) or NADPH (Apollo Scientific) and 100 mM of acetaldehyde (VWR), were used. For the inverse reaction, 0,1 M of Tris-HCl buffer (pH=8,8), 1 mM of NAD^+ (Sigma-Aldrich) or $NADP^+$ (Apollo Scientific) and 100 mM of ethanol, were used instead. Enzymatic assays were performed at 25°C in a total volume of 500 μL using a quartz cuvette (Hellma Analytics).

To calculate the blank for each reaction, 10 μL of the cofactor, 415 μL of the buffer and 25 μL of the cell-free extracts were added. In some cases, when the reaction was not detectable, the enzymatic assays were repeated using 50 μL of cell-free extracts instead. The OD_{340nm} was measured for 120 sec, with 10 sec of interval. To start the reaction, the substrate was added. For the direct reaction, 50 μL of acetaldehyde were subsequently added, and NADH/NADPH consumption was monitored by following the OD_{340nm} for 2 min. Regarding the inverse reaction, 50 μL of ethanol were added instead and NADH/NADPH formation was monitored.

2.5. Purification of Adh1 proteins of *St. bombicola* and *W. cacticola*

2.5.1. The *Escherichea coli* Rosetta constructs

The pET system was utilized to express the Adh1a from *St. bombicola* and the Adh1c from *W. cacticola*. The pET plasmids used in this experiment were pET19b (Amp+) and pET28a (Kan+). Both plasmids produce the repressor LacI that binds to its operator thereby regulating transcription of the gene to be expressed. The LacI repressor is removed from the operator when lactose (or a synthetic analogue such as IPTG) is added. These plasmids also encode a histidine tag that will be fused to the protein of interest, allowing the downstream separation and purification of the protein.

The *ADH1a* from *St. bombicola* was cloned in pET19b (pET19b:*ADH1*) and transformed in competent *E. coli* DH5α cells. The *ADH1c* from *W. cacticola* was cloned on the plasmid pET28a (pET28a:*ADH1*) in competent *E. coli* XLGold cells. Following the recovery of the constructed plasmid, *E. coli Rosetta* (DH3) was subsequently transformed with this final construct and used for the heterologous expression of Adh1 proteins. *E. coli Rosetta* is widely used to express eukaryotic proteins since it carries a plasmid that encodes tRNAs for codons that are rare in *E. coli* but common in eukaryotes (Rosano & Ceccarelli, 2014). This plasmid also confers resistance to chloramphenicol. The molecular constructs were obtained by Garvão, 2020.

2.5.2. Growth conditions

E. coli Rosetta (empty), *E. coli* pET19b:*ADH1* (*St. bombicola*) and *E. coli* pET28a:*ADH1* (*W. cacticola*) were pre-cultivated in 50 mL Erlenmeyer flasks, each containing 10 mL of LB (Luria broth) medium [1% (w/v) NaCl; 1% (w/v) Tryptone (BD Bacto); 0,5% (w/v) Yeast Extract]. For the *E. coli* Rosetta (empty) pre-inoculum, LB medium was supplemented with 34 μg/mL of chloramphenicol (Sigma-Aldrich). For the *E. coli* pET19b:*ADH1* (*St. bombicola*) pre-inoculum, LB medium was supplemented with 34 μg/mL of chloramphenicol and 100 μg/mL of ampicillin (NZYTech). For the *E. coli* pET28a:*ADH1* (*W. cacticola*) pre-inoculum, LB medium was supplemented with 34 μg/mL of chloramphenicol and 30 μg/mL of kanamycin (NZYTech). The *E. coli* Rosetta (empty) was used as a negative control in this experiment. The cells were incubated at 37°C overnight with orbital shaking.

2.5.3. Optimization of overexpression and solubility of Adh1 proteins

The production of an overexpressed and soluble protein is essential for protein purification. To test the optimal conditions for the soluble overexpression of Adh1 proteins of *W. cacticola* and *St. bombicola*, different parameters were tested on a small scale (using 250 mL Erlenmeyers, containing 50 mL of LB medium with the respective antibiotics). These parameters varied on IPTG concentrations (0,005 mM; 0,5 mM), incubation temperatures after induction (20°C; 25°C; 28°C; 37°C) and time after induction (T4; T8; T20).

For the expression optimization, cells from an overnight culture were inoculated in 250 mL Erlenmeyer flasks, containing 50 mL of the same medium, supplemented with the appropriate antibiotics. Optical densities were determined and cultures were diluted to an OD_{600}=0,06. The cells were incubated at 37°C for about 2 h 30 min with orbital shaking until they reached an OD_{600}~1. At this point, cultures were at T0. For each culture, optical densities were determined and four aliquots of 1 mL were collected to tubes that were subsequently centrifuged at 14.200 x g for 5 min at 4°C and pellets were stored at -20°C for SDS-PAGE (Sodium dodecyl sulphate-polyacrylamide gel electrophoresis) analysis. The induction of the expression of Adh1 was then performed by adding the defined concentration of IPTG (NZYTech) to each culture. Cells were incubated at the tested temperature until they reached induction time. Following this, the same procedure was repeated as T0. Four aliquots of 1 mL were collected to tubes, centrifuged and the pellets stored at -20°C, to confirm the overexpression by SDS-PAGE

analysis. Subsequently, 6 mL of the total 50 mL culture were centrifuged at 1.600 x g for 30 min at 4°C. The pellet was stored at -20°C until being used for cell lysis, to determine whether the evaluated condition resulted in a soluble protein.

Following the confirmation of overexpression and solubility by SDS-PAGE of a given condition, the experiment was reproduced on a large scale (using 2 L Erlenmeyers, containing 400 mL of LB medium, supplemented with the respective antibiotics). The experiment was performed as mentioned above and cells were collected at T0 and the defined hours after induction, for overexpression confirmation by SDS-PAGE. The cell culture was distributed to 50 mL falcons and centrifuged at 1.600 x g for 30 min at 4°C. The supernatants were discarded, and pellets were subsequently resuspended, using the remaining cell culture. A final centrifugation was performed on the resulting 50 mL falcon containing the pellet from all of the 400 mL culture, at 9.000 x g for 10 min at 4°C. The pellet was stored at -20°C for further cell lysis and subsequently protein purification.

2.5.4. Cell lysis through sonication

2.5.4.1. Cell lysis of cultures grown on small scale

While performing this experiment on a lower scale 1 mL of Working Lysis Phosphate Buffer Solution [20 mM Sodium Phosphate (pH=7,4); 500 mM Sodium Chloride; 0,6 mg Lysozyme (Muramidase from egg white, Sigma-Aldrich); 0,5 μL Benzonase nuclease (ChemCruz); 0,1 mM PMSF)] was prepared and 100 μL were added to the pellets that were collected from 6 mL of culture. The mixture was homogenized using the vortex and whenever the samples were too viscous (denoting the presence of nucleic acids), more Benzonase was added. Cell lysis through ultrasounds was performed with the Ultrasonic Processor UP200S (Hielscher), using the following parameters: 80% amplitude and 0,5 cycle. The samples were sonicated in 10 cycles of 1 min ultrasounds and 1 min on ice.

To understand whether the tested condition resulted in the overexpression of a soluble protein, 20 μL of the total extract were collected. The remaining extract was centrifuged at 9.300 x g for 10 min at 4°C and 20 μL of the supernatant were recovered. Both fractions were analysed by SDS-PAGE.

2.5.4.2. Cell lysis of cultures grown on large scale

While performing the experiment on a larger scale, 5 mL of Working Lysis Phosphate Buffer Solution were added to the harvested pellet from the 400 mL culture. This time, 3 mg of Lysozyme and 1 μL Benzonase nuclease were added instead. The mixture was further transferred to 15 mL falcons and homogenized using the vortex. Sonication was performed under the same parameters as mentioned above. After sonication, the overexpression of a soluble protein was confirmed by SDS-PAGE. Subsequently, 1 mL of the total extract was stored at -20°C to perform enzymatic assays and the remaining 4 mL were directly used for protein purification.

2.5.5. Adh1 purification

Protein purification was performed by immobilized metal ion affinity chromatography (IMAC), using 1 mL HisTrap HP columns (GE Healthcare). The IMAC columns are widely used for the purification of histidine-tagged recombinant proteins. Histidine efficiently binds to the transition metals (Co^{2+}, Ni^{2+}, Cu^{2+}, Zn^{2+}) that are immobilized on the column matrix. This allows the separation of the protein of interest from the remaining proteins of the extract. The elution of histidine-tagged proteins is subsequently performed by adding Imidazole. Imidazole is a histidine analogue that competitively binds to the column matrix, allowing for the detachment of the histidine-tagged proteins (Bornhorst & Falke, 2000).

To prepare the protein extract for purification, 20 mM of Imidazole (Sigma-Aldrich) were added. The low concentration of Imidazole is necessary to reduce unspecific binding of other proteins to the column (Bornhorst & Falke, 2000). The column was subsequently washed with 10 mL of filtered MiliQ water (Merck) and equilibrated with 10 mL of Binding Buffer [20 mM Sodium phosphate (pH=7,4); 500 mM Sodium Chloride]. The protein sample was charged onto the column and the flow-through (FT) was collected. The column was washed with 10 mL of the Binding Buffer containing 20 mM of Imidazole. Three washing fractions of 3 mL each, were collected (A1, A2, A3). FT, A1, A2, A3 were kept at -20°C until SDS-PAGE analysis.

To elute the protein of interest, 10 mL of the Elution Buffer [20 mM Sodium phosphate (pH=7,4); 500 mM Sodium Chloride], and 500 mM of Imidazole were added to the column. The eight elution fractions (1 mL each) were collected and analysed by SDS-PAGE gels. All the buffers, the MiliQ water and the sample were loaded into the column, using a 20 mL syringe (BBraun).

2.5.6. Sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE)

SDS-PAGE gels were performed in order to separate proteins from the *E. coli* Rosetta extracts and further evaluation of Adh1 overexpression. A 12% resolving gel was prepared with 40% acrylamide mix (acrylamide/bis-acrylamide 37.5:1, NZYTech), 1,5 M Tris-HCl (pH=8,8), 10% SDS (Sodium Dodecyl Sulphate, Sigma-Aldrich), 10% ammonium persulfate (APS, NZYTech) and tetramethylethylenediamine (TEMED, NZYTech). A 4% stacking gel was prepared with 40% acrylamide mix, 0,5 M Tris-HCl (pH=6,8), 10% SDS, 10% APS and TEMED.

The previously harvested cell pellets and the collected 20 μL of supernatants were used for SDS-PAGE analysis. The harvested pellets were re-suspended in a way that to an OD_{600nm}=1,2, 200 μL of 10% SDS are added. To prepare the samples for loading the SDS-PAGE gel, 5 μL of 5x SDS-PAGE sample loading buffer (NZYTech) were added to 20 μL of each sample. The mixtures were heated for 5 min at 95°C in a dry bath (Frilabo) and 10 μL of each mixture were loaded to the gel, as well as 3 μL of a protein marker (NZYColour Protein Marker II, NZYTech).

The electrophoresis was performed in TGS (Tris-Glycine-SDS) buffer (25 mM Tris; 192 mM glycine; 0,1% SDS) and the samples were separated by SDS-PAGE, using 1 mm mini-slab gels (Mini-PROTEAN II Electrophoresis System, Bio-Rad) at 150 V. For protein detection, the Coomassie R-250 Staining Solution (0,3% Coomassie Brilliant Blue R-250; 50% ethanol; 10% acetic acid) was added to the gel for 1 h. Following the removal of the Staining Solution, a de-staining solution (30% ethanol; 10% acetic acid) was subsequently added for 1 h. The gel was kept on a rocking platform (Biometra) while incubating with both solutions.

3. Results and Discussion

3.1. Phylogenomic analysis of the new W/S-clade genomes

The four distinct subgroups within the W/S clade were named according to the *ADH1* gene they encode. Species that have acquired the *ADH1a*, *ADH1b* or *ADH1c* type were placed on the subgroup A, B or C, respectively. There is also a fourth subgroup (*ADH0*) comprised by species that do not contain any *ADH1* gene in their genomes. This division was based on the available genomes at the time (Gonçalves et al., 2018, 2020). For this project, 12 new W/S-clade genomes were added to the previously 22 analysed genomes by Gonçalves et al., 2020 and a phylogenomic analysis was performed. This analysis was crucial to better elucidate the phylogenetic relationships between species and place the new genomes within the existent subgroups, as it is observed in Figure 3.1.

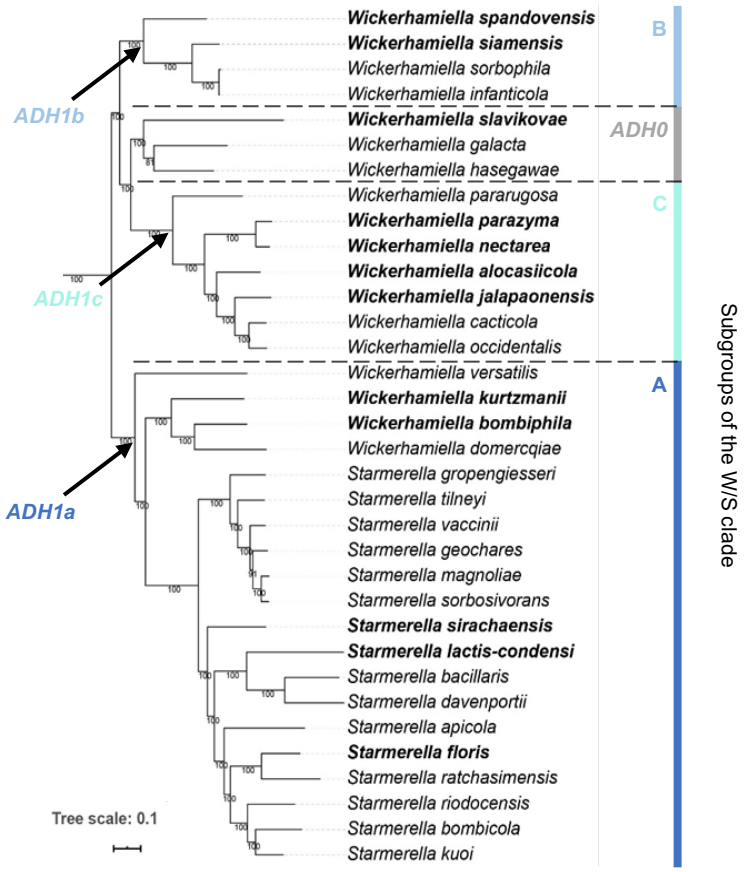


Figure 3.1. Phylogenomic analysis of W/S-clade species. The phylogenomic tree includes previously sequenced genomes (Gonçalves et al., 2020; Shen et al., 2018) and the 12 new genomes (highlighted in bold). The possible ancestor nodes where *ADH1a*, *ADH1b* and *ADH1c* were acquired are indicated with arrows. The distribution of species throughout the four inferred subgroups of the W/S clade (*ADH0*, A, B and C) is presented. Bootstrap support values (90-100) are indicated next to the respective nodes.

As a result of the phylogenomic analyses (Figure 3.1), it has been revealed that *W. bombiphila* and *W. kurtzmanii* are more closely related to the *Starmerella* genus than the other *Wickerhamiella* species,

similarly to what was previously observed for *W. domercqiae* and *W. versatilis* (Gonçalves et al., 2020; Shen et al., 2018). This further emphasizes that the *Wickerhamiella* genus is not monophyletic but paraphyletic.

Following the placement of the new genomes into the four subgroups (*ADH0*, A, B and C), the presence or absence of the alcoholic fermentation genes (*ADH1*, *ADH6*, *PDC/ARO10*) was subsequently evaluated in 26 W/S-clade genomes. This information was still completely obscure for the newly sequenced genomes. As for the remaining W/S-clade genomes that were evaluated in this project, only the presence or absence of *ADH1* had been confirmed (Gonçalves et al., 2020). This means the information regarding *ADH6* and *PDC/ARO10* also remained to be elucidated. The analyses of these genomes were crucial to better understand the dynamics of HGT in this clade, to identify the MRCA where the horizontal acquisitions took place and if each species has the *ADH1* type that is characteristic of its subgroup.

3.2. Identification of *ADH1* genes in W/S-clade genomes

3.2.1. Preliminary Adh1 phylogeny

To evaluate whether there were more HGT events involving *ADH1*, or other events such as duplications or gene losses, a phylogenetic tree of Adh1 proteins was reconstructed. The *Saccharomyces cerevisiae ADH1* gene sequence was used as a bait to search for orthologues in the 26 W/S-clade genomes. It was observed that for most of the analysed genomes, the *e*-value threshold used (*e*-value < $1e^{-5}$) was allowing for the selection of *ADH1* and *ADH6* sequences, because these two genes are closely related. To ascertain the orthology of the obtained BLAST hit sequences, a preliminary phylogenetic tree was constructed with the top 5.000 NCBI BLASTp hits, using the *Starmerella bombicola* Adh1 as query (Appendix; Figure A1). This step is important because Adh proteins from bacteria are not annotated as Adh1 and Adh6. Therefore, only by analysing the top hits, it is not possible to assure whether these sequences are orthologues.

By the examination of the preliminary Adh1 phylogeny (Appendix; Figure A1), it was possible to observe that the majority of sequences clustered with a group that belongs to the Adh1 family. The remaining sequences formed a separate group that is comprised by non-Adh1 sequences. These were subsequently removed from this analysis. Nevertheless, the sequences that were removed from the analysis were most likely Adh6 sequences since the top hits from the NCBI were NADPH-dependent Adh from bacteria.

3.2.1.1. Assessment of genomic contaminations

Contaminations are frequent problems in whole genome sequencing projects. The detection of these contaminations is crucial when inferring evolutionary events such as HGT and gene duplications. Detecting contaminations can be a difficult task but some information can be used to infer that a given sequence is most likely the product of contamination. This includes, for instance, sequences that are located in short scaffolds with low coverage (when compared to the average coverage of the entire genome). This is the case of a *W. hasegawae ADH1* sequence that was uncovered in the local BLAST search. The corresponding *ADH1* gene sequence was found in a low coverage (1,11765 x) and small contig (322 bp). When performing a BLASTp of this protein sequence against the NCBI database, the top hit showed a 100% identity with an Adh2 from *Candida auris*. Similar observations were obtained for a *W. occidentalis ADH1* sequence, found within a contig with 1,22957 x coverage and 328 bp, where the BLASTp searches showed an 86% similarity with *Candida arabinofermentans*.

To explore whether these sequences are the result of contaminations or possible pseudogenes, primers were designed for these regions and a PCR reaction was performed. No amplification was observed, confirming these sequences are probably the result of contamination during the sequencing process. A similar scenario was observed for *St. tilneyi* where two hits were found to be 73,02% and 88,48% similar to Adh2 sequences from *C. auris*, respectively. The length of these contigs was 356 bp and 497 bp and the coverage was 1,20266 x and 1,0543 x, respectively. In this case, confirmation by PCR was not possible because the culture was not available in the laboratory.

3.2.1.2. *W. slavikovae* (subgroup *ADH0*) and *W. jalapaonensis* (subgroup C) do not have an *ADH1*

The observation of the preliminary Adh1 phylogeny shows that all of the Adh-like sequences from *W. slavikovae* are clustered outside the Adh1 group (Appendix; Figure A1), which indicates that this species lacks Adh1-like sequences. Interestingly, *W. slavikovae* is placed on the same subgroup as the two *ADH0* species (*W. galacta* and *W. hasegawae*). This new evidence reinforces the idea that the MRCA of this subgroup did not horizontally acquire an *ADH1* gene, contrary to the remaining three subgroups. A similar result was obtained for *W. jalapaonensis*. In this case, this species is clustered within subgroup C. It seems that the acquisition of the *ADH1c* gene occurred in the MRCA of this subgroup, but it was subsequently lost in this species. The reason why it has only occurred for this species is not clear. To date, this is the only known W/S-clade species that has lost the HGT-acquired bacterial *ADH1* gene, constituting a case of secondary loss.

3.2.2. Three independent HGT events of bacterial *ADH1* to the W/S-clade

After selecting the putative Adh1 protein sequences from the preliminary tree, the top 4.000 BLASTp hits from Gonçalves et al., 2018 were added in order to reconstruct the final tree that is represented in Figure 3.2. By the observation of the final Adh1 phylogenetic tree, it is possible to conclude that all of the new W/S-clade sequences (pink) cluster with bacteria (blue) and not with other yeasts (red) or

filamentous fungi (orange). In this phylogenetic tree the new W/S-clade Adh1 sequences are placed according to the three previously identified subgroups (A, B and C). Each subgroup is clustered with different bacteria thus confirming the three independent HGT events, inferred by Gonçalves et al., 2020.

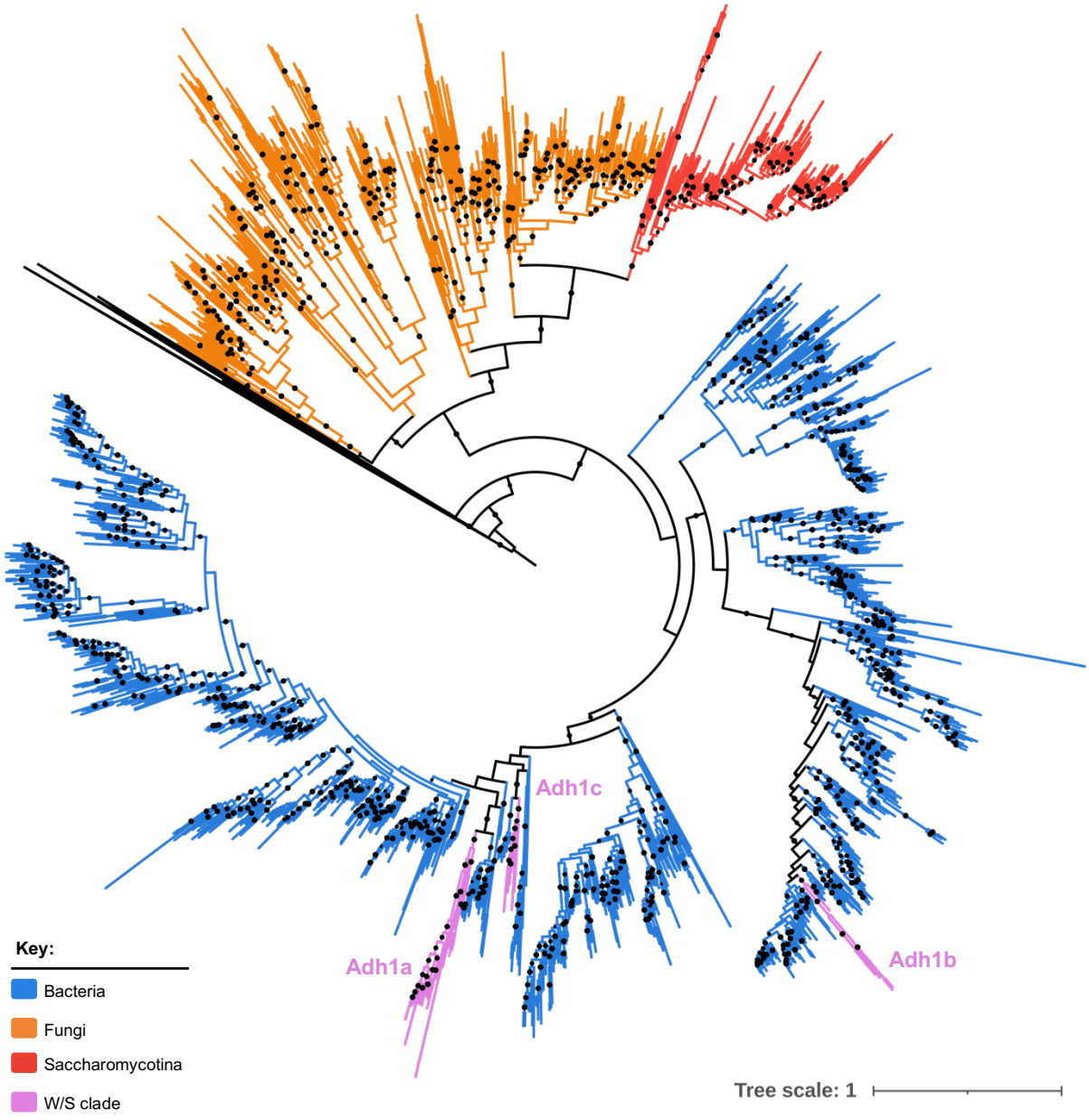


Figure 3.2. Reconstructed ML phylogeny of Adh1 proteins. The top 4.000 BLASTp hits from Gonçalves et al., 2018 were added to the selected Adh1 sequences from the preliminary tree (Appendix; Figure A1). All Adh1 sequences from the W/S clade, cluster with three different bacterial Adh1 groups, represented by Adh1a, Adh1b and Adh1c, confirming the three independent HGT events of bacterial *ADH1* to the W/S clade. The pruned trees of the Adh1a, Adh1b and Adh1c clusters are presented in Figure 3.3. Bootstrap values >90% are represented with black circles.

To better observe the phylogenetic relationships between species as well as the possible donor species of each subgroup, the Adh1 phylogenetic tree (Figure 3.2) was pruned and the three Adh1 clusters are represented in Figure 3.3. As it is possible to observe, the *ADH1a* gene that was inferred to belong to the *Starmerella* genus and two *Wickerhamiella* species (*W. domercqiae* and *W. versatilis*) (Figure 3.1), clusters with bacterial species from the Acetobacteraceae family (Figure 3.3 – A). All of the new

Starmerella genomes, that were previously placed within this subgroup, also harbour the *ADH1a* gene. The two new *Wickerhamiella* genomes (*W. bombiphila* and *W. kurtzmanii*), that were clustering with the subgroup A in the species tree were also confirmed to have a *ADH1a* type. This gives further support to the acquisition of *ADH1a* in the MRCA of *Starmerella, W. domercqiae* and *W. versatilis,* and also the fact that the *Wickerhamiella* genus is not monophyletic, but paraphyletic (Figure 3.1).

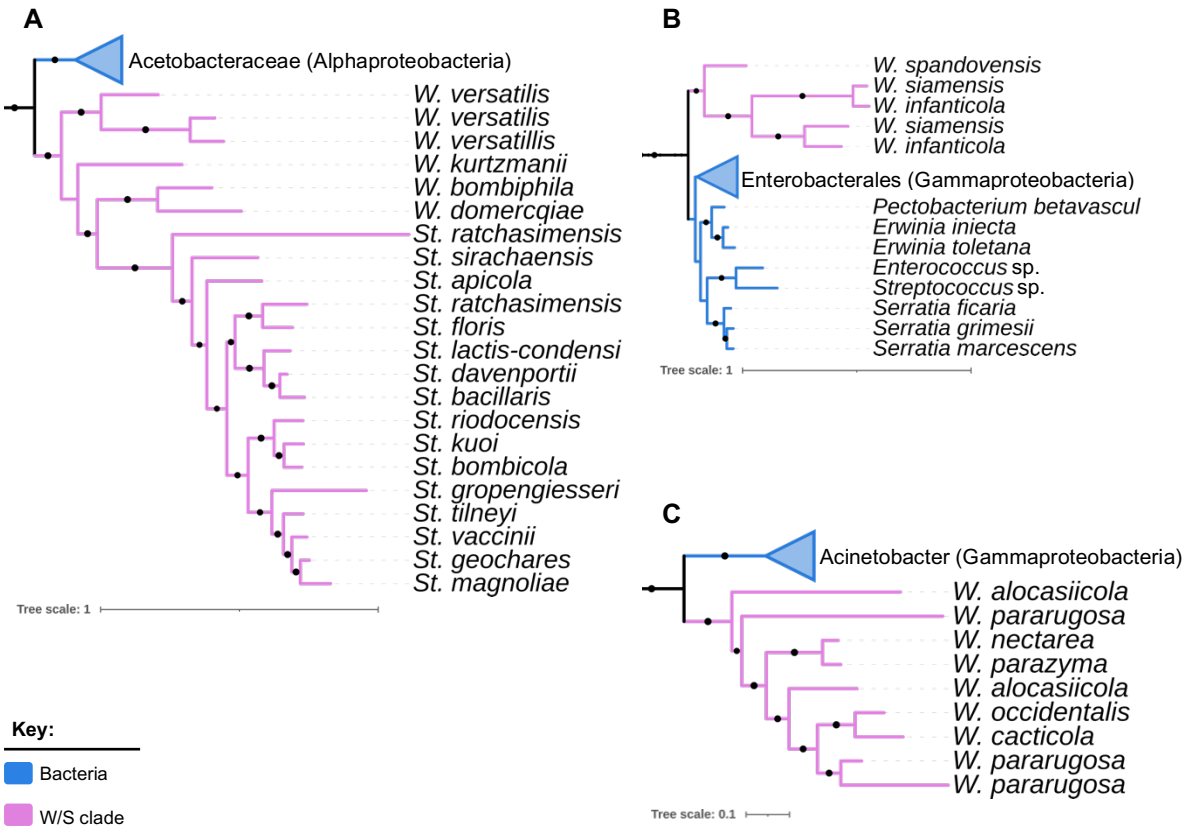


Figure 3.3. Pruned Adh1 phylogenetic trees of Adh1a, Adh1b and Adh1c clusters. The Adh1 phylogenetic tree from Figure 3.2 was pruned for the detailed observation of the groups containing different *ADH1* genes. (A) Adh1a sequences that cluster with the Acetobacteraceae (Alphaproteobacteria) (B) Adh1b sequences that cluster with the Enterobacterales (Gammaproteobacteria) and (C) Adh1c sequences that cluster with Acinetobacter (Gammaproteobacteria) sequences. Bootstrap values are represented by black circles (>90%).

Two intraspecific duplication events are observed in *W. versatilis*, which resulted in three paralogue genes in this species. The other species that has an *ADH1a* duplication is *St. ratchasimensis* with two paralogues. In this case, the duplication is possibly not species-specific, since the two sequences are not clustered together, but more sequences of closely-related species would be needed to put forward a hypothesis (Figure 3.3 – A).

Previous studies by Gonçalves et al., 2018 described that Adh1 sequences from *W. sorbophila* (that was previously classified as *Candida infanticola*) were probably acquired from the Firmicutes. However, the addition of the new genomes (*W. infanticola*, *W. siamensis* and *W. spandovensis*) clarified the phylogenetic relationship of Adh1b sequences, and it appears that the MRCA must have acquired the *ADH1b* gene from an Enterobacterales-related species instead (Figure 3.3 – B). As two Adh1

paralogues are present in *W. infanticola* and *W. siamensis* but not in *W. spandovensis*, it is possible to infer that a duplication event occurred in the ancestor of *W. infanticola* and *W. siamensis*.

Regarding the subgroup C, it seems that its MRCA acquired an *ADH1* gene from an Acinetobacter-related species (Figure 3.3 – C). In this case, paralogues were also detected in *W. alocasiicola* and *W. pararugosa*. Given the phylogenetic placement of the two paralogues of *W. alocasiicola*, it seems likely that a first duplication event occurred in the MRCA of subgroup C and then only one paralogue was maintained in most species. As for *W. pararugosa*, a second intraspecific duplication event occurred.

3.3. Identification of *ADH6* genes in W/S-clade genomes

3.3.1. *W. siamensis* (subgroup B), *W. hasegawae* (subgroup *ADH0*) and *W. spandovensis* (subgroup B) do not have an *ADH6*

The previously available genomes allowed the identification of one HGT event of a bacterial *ADH6* belonging to the Sphingomonadales (Alphaproteobacteria) (Gonçalves et al., 2018). In this project, the presence of *ADH6* within the 26 sequenced W/S-clade genomes was studied in order to look for more possible horizontal acquisitions, duplications or even losses of *ADH6*. The *S. cerevisiae ADH6* gene sequence was used to search for orthologues in the 26 W/S-clade genomes. Once again, a preliminary Adh6 phylogenetic tree was constructed to assess orthology. The top 5.000 BLASTp hits from NCBI (using one of the *St. bombicola* Adh6 paralogues as query) were used (Appendix; Figure A2).

The exclusion of sequences from the preliminary Adh6 phylogeny was based on the same logic as the one applied to the reconstruction of the Adh1 phylogeny (Figure 3.2). By removing the sequences that were clustered outside the Adh6 group, it was possible to infer that the sequences, preliminarily identified as *ADH6* in *W. siamensis*, *W. hasegawae* and *W. spandovensis* were not Adh6, which implies that these species lack an *ADH6*-like gene.

3.3.2. Three independent acquisitions of bacterial *ADH6* to the W/S clade

After selecting the Adh6 sequences from the preliminary tree (Appendix; Figure A2), the top 10.000 BLASTp hits from Gonçalves et al., 2018 were added to reconstruct the final tree that is represented in Figure 3.4. By the observation of the final Adh6 phylogenetic tree (Figure 3.4), it is possible to conclude that similarly to what was observed for the Adh1 phylogeny (Figure 3.2), the new W/S-clade Adh6 sequences (pink) cluster with bacteria (blue) and not with other yeasts (red) or filamentous fungi (orange). The available genomes at the time (Gonçalves et al., 2018, 2020) allowed the identification of a single HGT event that gave rise to the presence of *ADH6* in some W/S-clade species from subgroup

A. The addition of the new genomes has enlightened the patterns of horizontal acquisitions involving *ADH6* as two novel HGT events from bacteria were now unveiled (Figure 3.4).

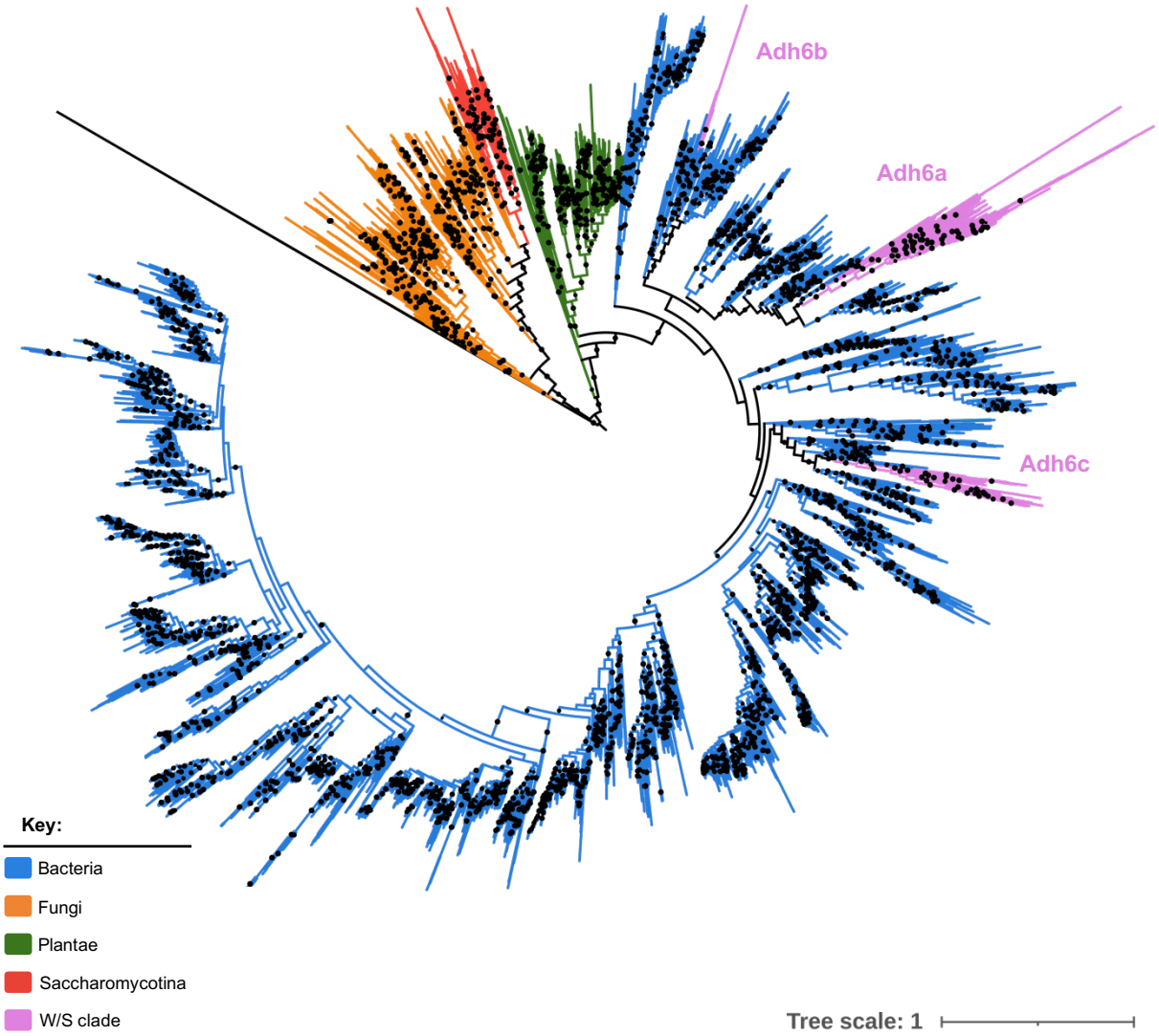


Figure 3.4. Reconstructed ML phylogeny of Adh6 proteins. The top 10.000 BLASTp hits from Gonçalves et al., 2018 were added to the selected Adh6 sequences from the preliminary tree (Appendix, Figure A2). All Adh6 sequences from the W/S clade, cluster with three different bacterial Adh6, revealing three independent HGT events to the W/S clade. The classification of Adh6a, Adh6b and Adh6c was based on the distribution the genes, to the respective W/S-clade subgroups. To clarify, the *ADH1a* and *ADH6a* genes belong to the subgroup A. The *ADH1c* and *ADH6c* genes belong to the subgroup C. As for *ADH6b*, it is exclusive of *W. slavikovae*. The pruned trees of the Adh6a, Adh6b and Adh6c clusters are presented in Figure 3.5. Bootstrap values are indicated with black circles (>90%).

To better observe the phylogenetic relationships between species as well as the possible donor species of each subgroup, the Adh6 phylogenetic tree (Figure 3.4) was pruned and the three Adh6 clusters are represented in Figure 3.5. It is clear that the *ADH6* genes have gone through many more duplication events than the *ADH1* genes. These events occurred either before or after speciation events.

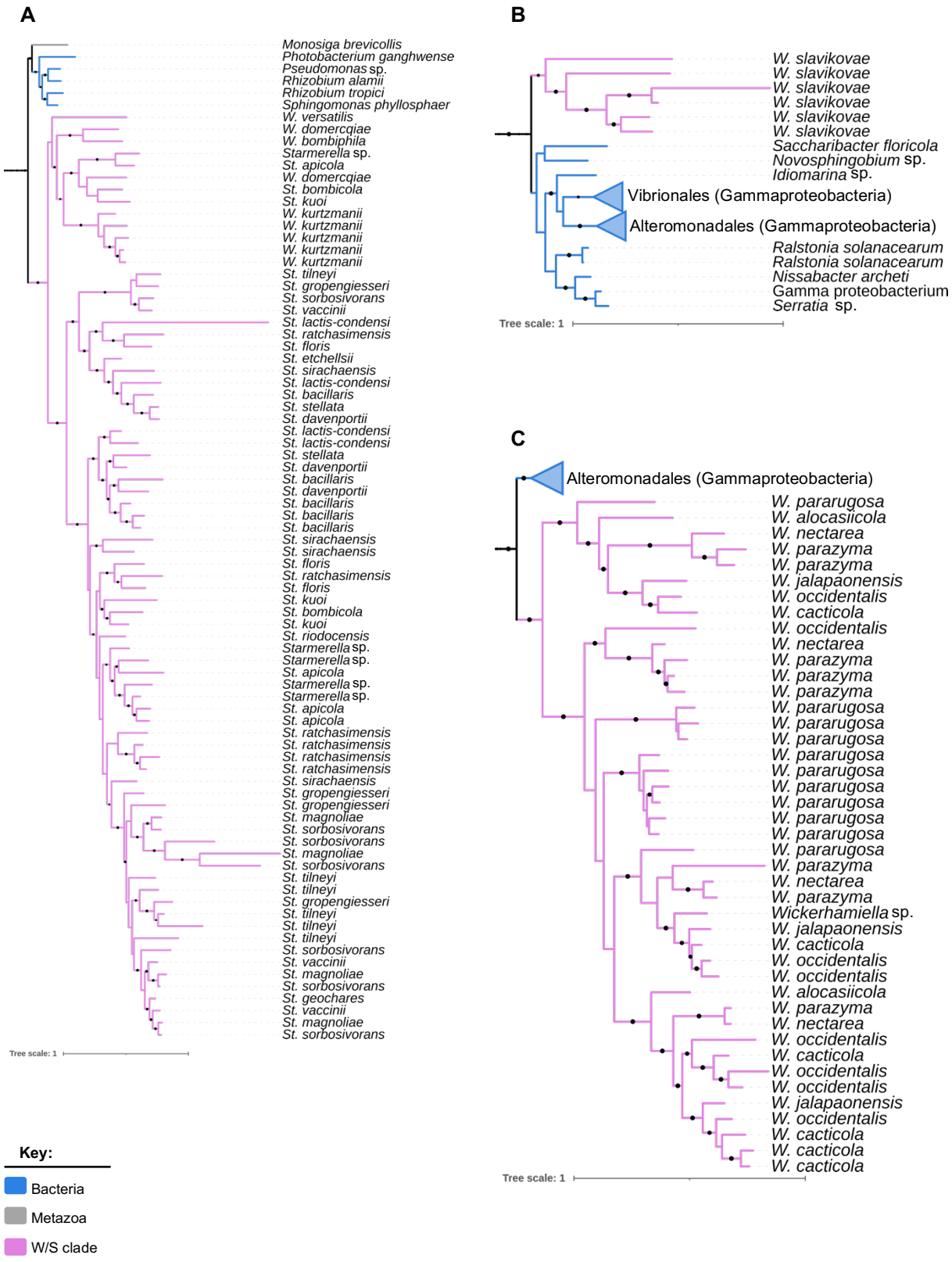


Figure 3.5. Pruned Adh6 phylogenetic trees of Adh6a, Adh6b and Adh6c clusters. The Adh6 phylogenetic tree from Figure 3.4 was pruned for the detailed observation of the groups containing different *ADH6* genes and their respective MRCA (A) Adh6a type with Sphingomonadales ancestry (Alphaproteobacteria) (B) Adh6b type from Vibrionales or Alteromonadales-related species (Gammaproteobacteria) (C) Adh6c type from Alteromonadales ancestry (Gammaproteobacteria). The Adh sequences that are identified as *Wickerhamiella* sp. or *Starmerella* sp. were obtained from W/S-clade genomes that have not been published yet. Bootstrap values are indicated with black circles (>90%).

The already reported HGT event involving the acquisition of the *ADH6* gene was inferred to have occurred from a Sphingomonadales-related species (Gonçalves et al., 2018, 2020). This HGT was confirmed to have occurred in the MRCA of all *Starmerella* and the *Wickerhamiella* species that cluster with *Starmerella* in the phylogenomic analysis (Figure 3.5 – A). Interestingly, this corresponds to the same ancestor where *ADH1a* was acquired. For this reason, the *ADH6* from the Sphingomonadales donor was named *ADH6a*.

The addition of the new genomes to the phylogenetic tree, showed that there were two additional HGT events that were previously unknown (Figure 3.5 – B, C). A second and newly observed HGT event was detected in the subgroup C (*W. alocasiicola*, *W. cacticola*, *W. jalapaonensis*, *W. nectarea*, *W. occidentalis*, *W. pararugosa* and *W. parazyma*) and was inferred that it involved an Alteromonadales related-species (Figure 3.5 – C). Curiously, except for *W. jalapaonensis* that seems to have lost its *ADH1c*, all the remaining species from this subgroup have the same *ADH1* gene. For that reason, this *ADH6* was named *ADH6c*.

The third acquisition of *ADH6* seems to be the first species-specific HGT event of a bacterial *ADH* gene to the W/S clade detected so far. Based on the current sampling, *W. slavikovae* is the only known species to have this bacterial *ADH6* (Figure 3.5 – B). The *ADH6a* and *ADH6c* were named based on the W/S-clade subgroups in which they were found. As for *W. slavikovae*, even though it belongs to the subgroup *ADH0*, this *ADH6* gene was named *ADH6b*.

Previous data (Gonçalves et al., 2018) described *W. galacta* as the only W/S-clade species that harbours a yeast-like *ADH6*. However, detection of *ADH6* in *W. galacta* might also be the result of a contamination since the sequence is present in a low coverage and small length contig. Regarding the species belonging to subgroup B (*W. infanticola*, *W. siamensis*, *W. spandovensis*), no *ADH6*-like sequences were found.

3.4. Identification of *PDC1*/*ARO10* genes in W/S-clade genomes

The loss of alcoholic fermentation was previously hypothesized based on the absence of native *ADH1* and *PDC1* genes. The reinstatement of the alcoholic fermentation pathway in some W/S-clade species was made possible through the acquisition of *ADH1* genes from bacteria and the co-option of a pre-existing decarboxylase, Aro10. This decarboxylase is phylogenetically very closely related to Pdc (Romagnoli et al., 2012).

Similarly to the methodology followed for Adh proteins, identification of putative *PDC* genes was also performed by tBLASTx searches in W/S-clade genomes, using the *S. cerevisiae PDC1* sequence as a bait. Pdc proteins are phylogenetically closely related to Aro10 proteins, therefore it is envisioned that significant blast hits (*e*-value < $1e^{-5}$) are obtained. An exception to this was *W. slavikovae*, for which no significant hit was obtained using this *e*-value threshold, which indicates that both Pdc and Aro10 sequences are absent. The *e*-value is influenced by the size of the database and also by the length of

the sequences. To confirm that this result was not due to fragmentation of the genome (which can result in fragmented gene sequences), the best obtained hit (ignoring the *e*-value threshold) was blasted against the NCBI database. The sequence presented a 75,87% identity to an Acetolactate synthase from *W. sorbophila* (*C. infanticola*), therefore supporting that both *PDC*- and *ARO10*-like genes are absent in *W. slavikovae*. Curiously, *W. slavikovae* does not carry an *ADH1* gene. However, it has several copies of the HGT-derived *ADH6* (Figure 3.5 – B).

In the remaining W/S-clade genomes, significant hits were obtained in the tBLASTx searches. Hence, to observe if these sequences were Pdc or Aro10, a phylogenetic tree was subsequently reconstructed. This phylogenetic tree was constructed by using the predicted Pdc/Aro10 protein sequences along with the sequences used in Gonçalves et al., 2018. This phylogeny is represented in Figure 3.6.

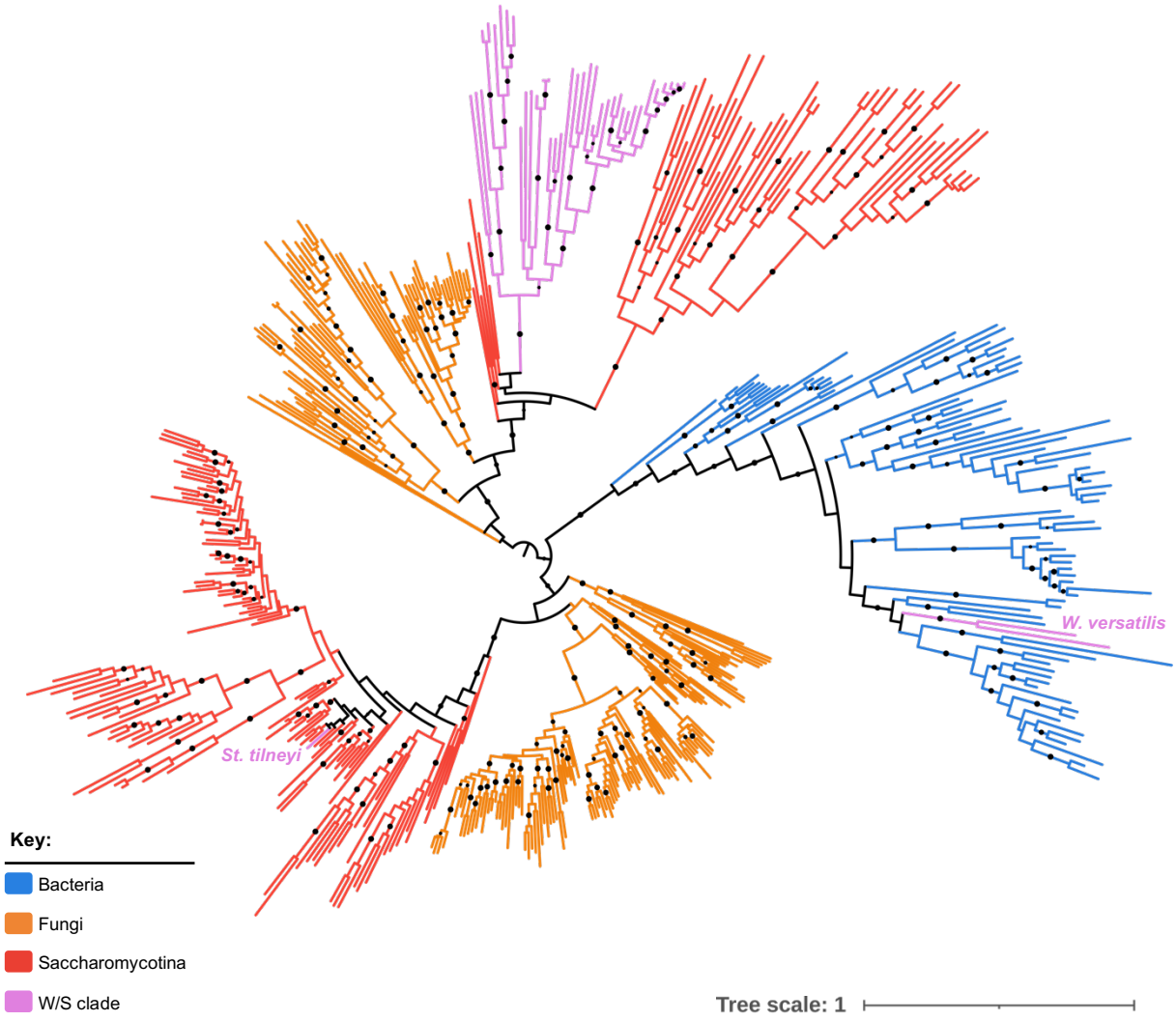


Figure 3.6. Reconstructed ML phylogeny of Pdc1/Aro10 proteins. The Pdc-like and Aro10-like sequences from Gonçalves et al., 2018 were added to the Pdc1/Aro10 predicted protein sequences from the 26 genomes. The pruned tree of the Aro10 cluster is presented in Figure 3.7. Bootstrap values are indicated with black circles (>90%).

As it possible to observe, the majority of Pdc-like sequences cluster with the fungal Aro10. Similarly to what was observed for *St. bombicola* (Gonçalves et al., 2018), it is possible that Aro10 took over the

role of the Pdc1 enzyme in these W/S-clade species, if they are capable of conducting alcoholic fermentation. The only known exception to this is *W. versatilis*, which seems to have acquired a *PDC1* from an Actinobacteria-related species (Gonçalves et al., 2018). A partial Pdc sequence was found in *St. tilneyi* and it is clustered with other Pdc1 sequences from other yeasts. However, similarly to what happened with Adh sequences, it is found in a low coverage contig (522 bp contig with a 1,02355 x coverage) that has a 92,98% identity with a pyruvate decarboxylase from *Metschnikowia bicuspitada*. It is possible to conclude that this genome has several contigs that possibly resulted from contamination issues during the sequencing project. In general, small contigs (<1.000 bp) with low coverage are not taken into account for genomic analyses, but in this case, they were not excluded so possible interesting data would not be missed.

In order to better observe the phylogenetic relationships between the Aro10 proteins, the phylogenetic tree from Figure 3.6 was pruned and is observed in Figure 3.7. As it is possible observe, the *ARO10* gene has duplicated several times in some species. The observed duplications are also before and after speciation events. Even though this phenomenon was found to be common among alcohol dehydrogenase genes (especially *ADH6*), this is the first time that it is observed on *ARO10* from the W/S clade.

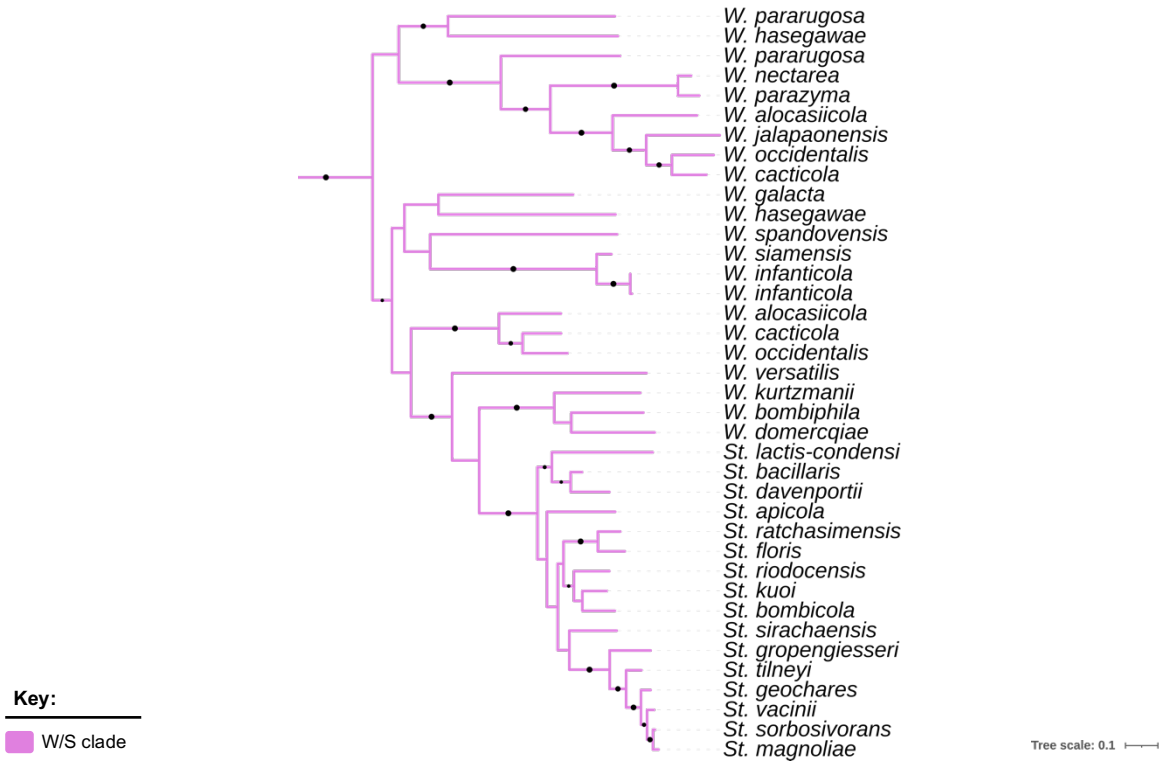


Figure 3.7. Pruned Aro10 tree of W/S-clade sequences. The Aro10 phylogenetic tree from Figure 3.6 was pruned for the detailed observation of the phylogenetic relationships between species. Bootstrap values are indicated with black circles (>90%).

3.5. Updated phylogenetic relationships and dynamics of HGT in the W/S clade

After collecting all data from the 26 sequenced genomes, it was possible to better elucidate the phylogenetic relationships and patterns of HGT of *ADH1*, *ADH6* and *PDC1*/ARO10 genes. A scheme summarizing the data regarding the new and the previous information (Gonçalves et al., 2018, Gonçalves et al., 2020) is presented in Figure 3.8. It is now possible to infer that both *ADH1* and *ADH6* genes underwent at least three independent HGT events each, possibly from different bacterial donors. The native *PDC1* gene is absent in all genomes but *ARO10* is present in all species except for *W. slavikovae*.

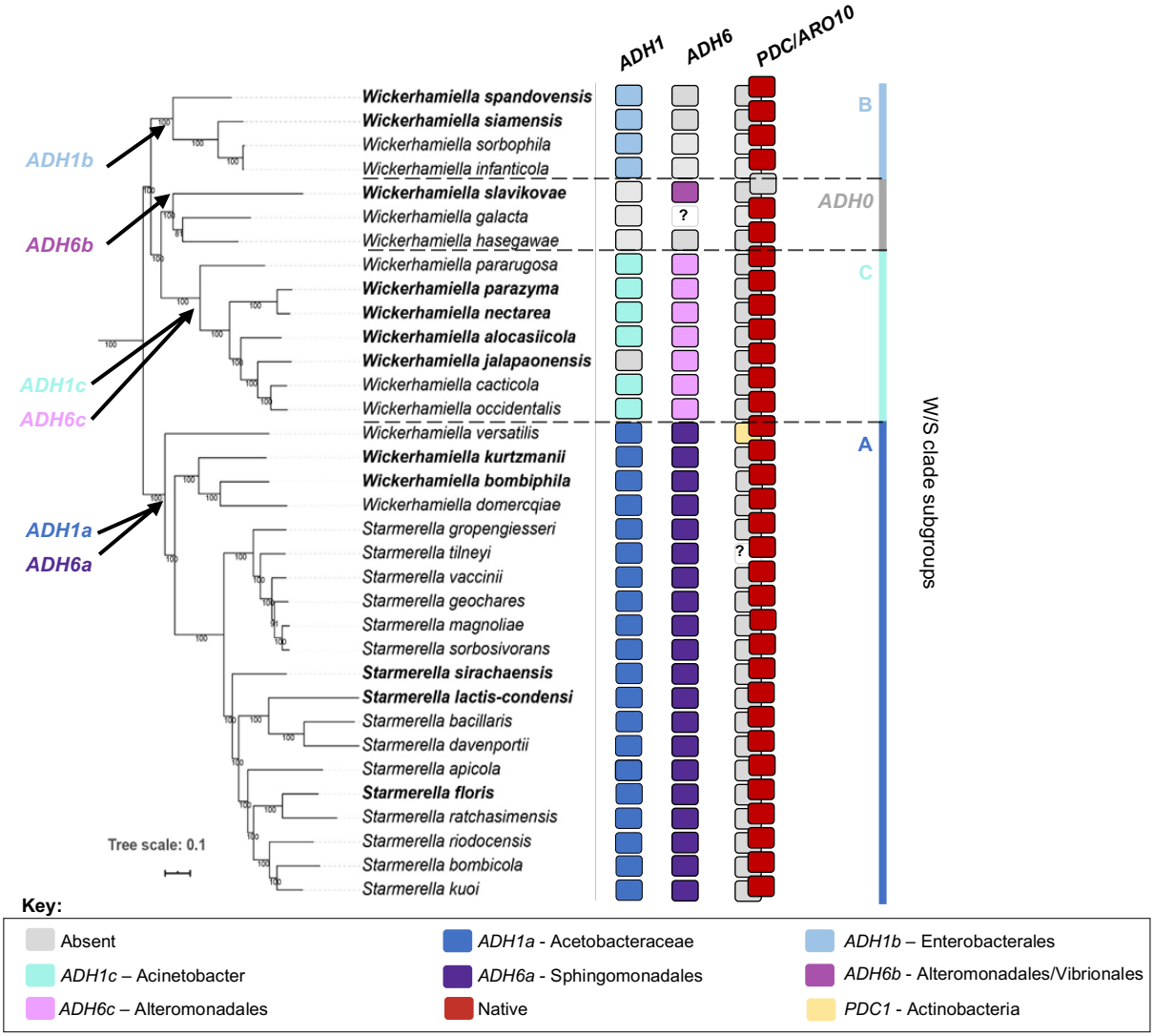


Figure 3.8. Phylogenetic relationships and dynamics of HGT in the W/S clade, including new data. Maximum likelihood (ML) phylogenomic tree and distribution of *ADH1 (ADH1a*, *ADH1b*, *ADH1c*), *ADH6* (*ADH6a*, *ADH6b* and *ADH6c*) and *ARO10/PDC1* genes in all available genomes of the W/S clade. The putative donor species are represented by colours, as indicated in the key. The postulated HGT events are indicated by arrows. Bootstrap support values (90-100) are indicated next to the respective nodes. The W/S-clade species presented are distributed within four distinct subgroups (*ADH0*, A, B and C) according to the *ADH1* gene they own (or the absence of the gene, in the case of the subgroup *ADH0*). In *W. galacta* and *St. tilneyi* the *ADH6* and the *PDC* gene, respectively, are indicated with a "?" since it was not possible to determine whether the predicted sequences are a contamination or not.

Taking into account the available genomic data, it seems that *ADH1* and *ADH6* genes were acquired at a not-too-distant point in evolution. For instance, the MRCA of *Starmerella*, *W. domercqiae* and *W. versatilis* acquired an *ADH1a* gene from Acetobacteraceae and an *ADH6* (*ADH6a*) gene from Sphingomonadales. It is possible to take a similar conclusion regarding the MRCA of the species containing the *ADH1c*. While the *ADH1c* gene was acquired from the Acinetobacter, the *ADH6*c gene was acquired from the Alteromonadales. Even though *W. jalapaonensis* does not have any *ADH1* (being the exception in subgroup C), it has the same *ADH6* as this subgroup (*ADH6c*). This evidence further emphasizes that *ADH1c* and *ADH6c* were acquired from different donors in the MRCA of this subclade and *ADH1c* was only lost in *W. jalapaonensis.* The period of time between the acquisitions of *ADH1* and *ADH6* must have been relatively short, given the consistent presence of the same *ADH1* and *ADH6* between subgroups. Regarding the subgroups B and *ADH0* (excluding *W. slavikovae*), it seems that both did not horizontally acquire an *ADH6* gene.

The results obtained by the study of the alcoholic fermentation genes in additional W/S-clade genomes, increased the robustness of the phylogenomic analysis of this clade (Figure 3.8). The distribution of species throughout the four subgroups is also supported by the presence or absence of the respective *ADH1* and *ADH6* bacterial genes. It is now possible to infer with more certainty the internal nodes where the acquisition of *ADH-like* genes took place. Also, two new *ADH6* acquisitions (*ADH6b* and *ADH6c*) were described.

3.5.1. The new data support the '*loss followed by reacquisition*' hypothesis

This analysis also gave more support to the '*loss followed by reacquisition*' hypothesis, firstly described by Gonçalves et al., 2018. Following the evaluation of the 26 genomes, there are still no species that have a native *ADH1* or *PDC1* genes. The absence or the sole presence of HGT-derived *ADH1*, highlight the fact that these were acquired after the total loss of the alcoholic fermentation pathway. The *W. slavikovae* species, that was added to the subgroup *ADH0*, also lacks the *ADH1* gene, giving further support that there is a lineage where the horizontal acquisition of *ADH1* did not take place. With the exception of *W. slavikovae*, most species have a *ARO10* that is possibly fulfilling the role of the lost *PDC1*. The loss of both *ADH* and *PDC* genes must have occurred in a relatively quick succession since no analysed genome has fungal *PDC* or *ADH* genes.

3.6. Alcoholic fermentation in W/S-clade species

3.6.1. Ethanol production and assimilation profiles in W/S-clade species

The *PDC1* and *ADH1* genes encode the enzymes responsible for conversion of pyruvate (from glycolysis) into acetaldehyde, and from acetaldehyde into ethanol, respectively. The presence of bacterial *ADH1* and *ADH6* genes, along with the hypothesized replacement of *PDC1* by *ARO10* in the majority of the W/S-clade genomes evaluated, indicate that these yeasts were most likely able to re-establish their fermentative capacities.

The Adh1a enzyme from *St. bombicola* was the only one whose involvement in alcoholic fermentation and growth on ethanol as the sole carbon source was tested (Gonçalves et al., 2018). It is still not clear if Adh1b and Adh1c enzymes also play a role in alcoholic fermentation. However, since the *ADH1b* and *ADH1c* genes were horizontally acquired and maintained in most analysed genomes (with the exception of *W. jalapaonensis*), they most likely play an important role in metabolism. Otherwise, there was a greater probability that the genes would be lost or would become pseudogenes (Fitzpatrick, 2011; Lindsey & Newton, 2019).

W/S-clade yeasts have also horizontally acquired distinct *ADH6* genes from different bacterial species. Even though Adh6 and its paralogue Adh7 are not the main enzymes involved in alcoholic fermentation in *S. cerevisiae*, they can participate in alcoholic fermentation when the other Adh enzymes are eliminated (de Smidt et al., 2008, 2011). As for the W/S clade, the elimination of the two *ADH6a* paralogues of *St. bombicola* does not seem to impair ethanol production. It is likely that in this clade, *ADH6* does not have a central role in the alcoholic fermentation pathway but is rather involved in other metabolic processes (Gonçalves et al., 2018). Regarding the two new bacterial *ADH6* genes uncovered in this work (*ADH6b* and *ADH6c*), their involvement in alcoholic fermentation remains to be evaluated.

It is crucial to complement *in silico* data concerning the distribution and origins of *ADH1*, *ADH6* and *PDC1/ARO10* genes amongst W/S-clade species with physiological data. This allows to elucidate whether the respective enzymes confer different phenotypic traits to the W/S-clade species that carry them. To achieve that, a subset of well distributed species with representatives from all subgroups, was selected to evaluate ethanol production and assimilation, and the results are presented in Table 3.1.

The majority of the published data was retrieved from the CBS culture collection database (Westerdijk Fungal Biodiversity Institute). Whenever this information was not available in the CBS database, it was retrieved from the Portuguese Yeast Culture Collection (PYCC) database or the research papers where the species were initially described. Most of the obtained experimental results are in line with both the databases (CBS and PYCC) and the literature. The observed discrepancies can be explained by several factors. For example, in this experiment ethanol assimilation was tested in YNB supplemented with 2% ethanol. YNB is a minimal medium in which some species may not be able to grow vigorously. Besides this, there is always an inherent intraspecific variation. These stochastic variations greatly influence the observed phenotypes. Even when it is possible to reduce the genetic and environmental variations to the maximum, phenotypic variability is still observable (Daniel & Meyer, 2013; Geiler-Samerotte et al., 2013; Yvert et al., 2013).

Table 3.1. Detection of alcoholic fermentation and ethanol production in W/S-clade species. Ethanol production and assimilation of several species from the different W/S-clade subgroups (*ADH0*, A, B and C). Ethanol production was assessed through HPLC and was considered positive (+) when the measured concentration was higher than 5,00 g/L. When ethanol concentration was lower than 5,00 g/L, the respective maximum concentration in g/L is indicated. Ethanol assimilation was considered based on the macroscopic observation of growth in minimal medium YNB + 2.0% ethanol. Ethanol assimilation was considered 'weak' or 'delayed' when the culture did not grow vigorously in ethanol or it took more than 15 days for the growth to be observable, respectively. Experimental results are compared, with physiological information, retrieved from the literature. The species *W. jalapaonensis* is marked with a * since it belongs to the *ADH1c* phylogenetic subgroup but lost the gene. The species *W. vanderwaltii* and *W. kazuoi* are marked with a [&] since the placement of these two species, on the respective subgroups, was made according to phylogenies based on the D1/D2 phylogeny (de Vega et al., 2017). It is still not confirmed whether they harbour *ADH* genes and their phylogenetic placement.

	Ethanol production		**Ethanol assimilation**	
Subgroup *ADH0*	Experiment	Published data	Experiment	Published data
W. galacta	-	-	weak	-
W. hasegawae	4,16	weak/delayed	-	-
W. kazuoi[&]	-	-	-	-
Subgroup A				
St. bacillaris	+	+	-	-
St. apicola	+	+	-	-
W. domercqiae	0,21	-	-	+
Subgroup B				
W. infanticola	-	-	-	+
W. siamensis	0,04	-	+	weak
Subgroup C				
W. cacticola	-	-	+	weak
W. occidentalis	-	-	+	+
W. parazyma	-	-	+	weak/delayed
W. pararugosa	0,06	-	+	+
W. nectarea	-	-	weak	weak/delayed
W. vanderwaltii[&]	2,93	-	+	+
W. jalapaonensis[*]	-	-	-	weak

3.6.1.1. Ethanol production and assimilation in the subgroup *ADH0*

With the exception of *W. hasegawae,* the remaining *ADH0* species did not produce any ethanol. This result is in line with the information retrieved from the CBS database, which referred that *W. hasegawae* presents a weak and delayed ethanol production (fermentation). This is a very curious result given that no *ADH1* or *ADH6* genes were found in the *W. hasegawae* genome (Figure 3.8).

The ethanol trace that is observed may be due to the presence of a secondary enzyme also involved in alcoholic fermentation, for instance Sfa1. The presence of this enzyme was detected in all W/S-clade genomes evaluated in Gonçalves et al., 2018. Despite the fact that Adh1 is the main enzyme for ethanol production in *S. cerevisiae*, other alcohol dehydrogenases are capable of engaging in alcoholic fermentation when Adh1 is absent (de Smidt, 2008, 2011; Drewke, 1990; Ida, 2011). The same could be occurring in *W. hasegawae.* Since the only Adh present is Sfa1, it could happen that the enzyme is partially producing trace amounts of ethanol, in the absence of both Adh1 and Adh6.

Regarding ethanol assimilation, with the exception of *W. galacta*, the remaining *ADH0* species were not able to grow on ethanol as the sole carbon source. It is still not clear how *W. galacta* is able to weakly grow on ethanol. A possible explanation could be the Sfa1 enzyme, as it was mentioned for *W. hasegawae*. Nevertheless, none of these species show strong ethanol production or assimilation profiles, which is line with the genomics data that indicate the absence of the main Adh-like enzymes.

3.6.1.2. Ethanol production and assimilation in the subgroups A, B and C

Contrary to what was observed for *St. bombicola*, that is able to produce and assimilate ethanol (Gonçalves et al., 2018), the remaining species of the subgroup A do not show the same profile. The species *St. bacillaris* and *St. apicola* are able to produce significant amounts of ethanol but ethanol assimilation was not observed. Moreover, *W. domercqiae,* which also carries an Adh1a type enzyme, is not able to produce nor assimilate ethanol. However, when cross-referencing this information with the database, assimilation of ethanol was described for this species.

The selected species from the subgroup B (*W. infanticola* and *W. siamensis*) are only able to assimilate ethanol. The species from the subgroup C also show a clear tendency for the assimilation of ethanol in detriment of fermentation. These results point to the possibility that the Adh1b and Adh1c enzymes are mediating, preferentially, the conversion of ethanol into acetaldehyde, at least under the conditions tested. In line with the absence of any Adh1 enzyme, *W. jalapaonensis* is not able to assimilate nor produce ethanol. However, the published data indicates that this species is able to weakly assimilate ethanol, which in this case, can be explained by the presence of Adh6 or Sfa1.

These results show that, independently of the origin of the Adh enzymes, there is significant variability. The different production/assimilation profiles might indicate that each species is either regulating the enzymes differently or these have evolved to have different specificities. Different species may be producing Adh proteins with different conformations, which can facilitate the direct or the inverse reaction (and the binding of NAD(P)H or $NAD(P)^+$, respectively). Nevertheless, it is possible to observe that at least one representative of each of the subgroups A, B and C has the ability to substantially produce and/or grow on ethanol, indicating that some Adh enzymes might be functional.

3.6.2. Cofactor preference

The ability to produce/assimilate ethanol indicates that Adh1 enzymes might be functional, however, to better understand their involvement in alcoholic fermentation, enzymatic assays using total protein extracts were performed. The assessment of enzymatic activities can also shed some light on the cofactor preferences of the different Adh1 enzymes. To achieve this, eight species from the W/S clade representing each subgroup were selected. Two non-W/S-clade species were also chosen for comparison: one yeast species that is closely related to the W/S clade (*Candida incommunis*) (Shen et al., 2018) and a possible bacterial donor of *ADH1a* (*Acetobacter malorum*). The results are summarized on Table 3.2.

Table 3.2. Relative alcohol dehydrogenase activities of total protein extracts of W/S-clade and non-W/S-clade species. The relative alcohol dehydrogenase activities of total protein extracts were measured using NADH, NADPH, NAD^+ and $NADP^+$ as cofactors. For each species, relative enzymatic activities were calculated in relation to the strongest detected reaction (indicated with 1,000). Absence of any measurable activity is indicated with "x". The enzymatic activities were tested on representatives of each W/S-clade species subgroups (*ADH0*, A, B and C). The bacterial species *A. malorum* is a possible donor of *ADH1a* to the W/S clade and *C. incommunis* is a yeast species that is closely related to the W/S clade. The *W. jalapaonensis* species is marked with a * since it belongs to the subgroup C but lost the Adh1c gene. *W. vanderwaltii* and *W. kazuoi* are marked with a & since the placement of these two species, on the respective subgroups, was made according to phylogenies based on D1/D2 domain, of the large subunit rRNA gene from de Vega et al., 2017. It is still not confirmed whether they have *ADH* genes and their phylogenetic origin. Additionally, *W. infanticola* and *S. cerevisiae* were added to the results (Diamantino, 2020). While *W. infanticola* is a representative of the subgroup B, *S. cerevisiae* is added for comparison, since it is the yeast species with the most well characterized Adh enzymes.

	Cofactors			
Subgroup *ADH0*	**NADH**	**NADPH**	**NAD^+**	**$NADP^+$**
W. galacta	x	x	x	1,000
W. hasegawae	0,830	0,370	1,000	x
W. kazuoi&	x	1,000	x	x
Subgroup A				
St. bacillaris	x	1,000	x	x
St. apicola	1,000	0,080	x	x
W. domercqiae	x	x	x	x
Subgroup B				
W. infanticola	Activity	Activity	Activity	x
Subgroup C				
W. vanderwaltii&	0,640	1,000	x	x
*W. jalapaonensis**	0,001	x	1,000	x
Non-W/S-clade species				
A. malorum	0,110	x	1,000	0,020
C. incommunis	0,780	0,070	1,000	x
S. cerevisiae	Activity	x	Activity	x

Candida incommunis was selected because it harbours native *ADH1* genes and is phylogenetically closely related to the W/S clade (Shen et al., 2018). It is important to assess the enzymatic activities of this species because it is described that yeasts use NADH and not NADPH during alcoholic fermentation, while it was shown that Adh1a from *St. bombicola* is probably able to use both cofactors (Gonçalves et al., 2018). This will contribute to understand if the use of NADPH constitutes an innovation of the W/S Adh1 enzymes. A representative of the possible donor lineage (*A. malorum*) was selected to elucidate whether the use of NADPH is a characteristic of the horizontally acquired enzymes or might have evolved post acquisition in the yeast host.

It is important to take into consideration that enzymatic assays of total protein extracts do not distinguish between Adh1, Adh6 and Sfa1 activities. The extract contains all the proteins that were being expressed at the moment of cell harvest. Moreover, enzyme inhibitors might also be present. However, it was expected that the majority of the measured enzymatic activities would come from the Adh1-like proteins since they display the highest affinity for ethanol/acetaldehyde (de Smidt et al., 2011).

Generally, alcohol dehydrogenase activities were detected in all species with the exception of *W. domercqiae*. Despite not being able to detect any Adh1-like or Adh6-like enzymes in *W. galacta* and *W. hasegawae*, alcohol dehydrogenase activity was detected, and this could be explained by the presence of other unspecific dehydrogenases. As for *W. kazuoi*, its genome is not available but taking into account the D1/D2 phylogeny, one can assume that it belongs to the subgroup *ADH0*, therefore lacking a Adh1-like enzyme. To shed light into this, genomic analysis of this species is essential.

The results for species belonging to subgroup A are in line with the physiological data (Table 3.1). As for the cofactor specificity, *St. apicola* is able to use both NADH and NADPH as cofactors, displaying a preference for NADH (activity is ~13x higher), which is in agreement with previous results obtained in Gonçalves et al., 2018. However, for *St. bacillaris* no activity was detected when NADH was used as a cofactor, contrarily to what was observed in Gonçalves et al., 2018. This can be either explained by experimental factors and the inherent variability while measuring enzymatic activities.

Alcohol dehydrogenase activities in *W. infanticola* (subgroup B) and *W. vanderwaltii* (subgroup C) were detected using both NADH and NADPH as cofactors. In both species, there is a slight preference for the use of NADPH over NADH. The species *W. vanderwaltii* is able to grow on ethanol (Table 3.1) but does not present any Adh activity on the inverse reaction. This emphasizes the hypothesis that inhibitory effects might be present. The detection of Adh activity in the inverse reaction was experimentally difficult. In Gonçalves et al., 2018, only the direct reaction (using both NADH and NADPH) was detected in *St. bombicola*, however this species is able to produce and assimilate ethanol (Gonçalves et al., 2018).

3.6.2.1. Cofactor preference on non-W/S-clade species

The putative donor of *ADH1a* belongs to the Acetobacteraceae family. Characterization of Adh enzymes from *Acetobacter pasteurianus* indicate that this enzyme is NADH-dependent (Masud et al., 2011). However, it was not clear whether the NADPH cofactor was tested. In the current experiment, a different Acetobacteraceae species was used, *Acetobacter malorum*. It was possible to detect Adh activity using NADH (low) and NAD^+ and negligible activity with $NADP^+$ but not with NADPH. This might indicate that Acetobacteraceae Adh preferentially uses NAD(H) and suggests that the ability of the Adh1a enzyme to use NADP(H) could have evolved in the host yeast species.

The yeast *S. cerevisiae* has the most well characterised Adh proteins. As it was previously studied, the Adh1-like activity from *S. cerevisiae* is NAD(H)-dependent (de Smidt, 2008, 2011). No NAD(P)-dependent activity was detected. The yeast *C. incommunis* is a closely related species to the W/S clade whose genome is available, and it is confirmed that this species owns a yeast native Adh (Gonçalves et al., 2020; Shen et al., 2018). According to the results, this species is able to use NAD^+. Adh activity was detected in *C. incommunis* extracts using both NADH and NADPH as cofactors, however the activity with NADH is ~11x higher when compared to the activity with NADPH. This may indicate that native yeast Adh1 enzymes can also accept NADPH as cofactors. To confirm this, more species outside the W/S clade must be tested. Moreover, activities using cell-free extracts can only provide a preliminary

idea about the overall involvement of Adh enzymes in alcoholic fermentation, but it is not possible to discriminate between the different Adh types (Adh1-like, Adh6-like or Sfa1-like). To accurately characterize the activities of Adh1 enzymes, enzymatic assays using pure proteins is the most appropriate method.

3.7. Purification and characterization of Adh1 enzymes from *St. bombicola* and *W. cacticola*

3.7.1. Overexpression of Adh1 proteins

Enzymatic characterization is more accurate while working with purified proteins when compared to total protein extracts, which contain all proteins that are being translated at the moment of culture harvest. This means that parallel reactions can occur when testing a given enzymatic activity and enzyme inhibitors can also interfere with the assay. In order to more accurately characterize Adh1 activities, the Adh1a from *St. bombicola* and the Adh1c from *W. cacticola* were chosen for purification. The Adh1 from these two species was selected since they do not carry any other *ADH1* paralogue.

The process of protein purification often leads to a decrease in concentration, so it is important to overexpress the protein of interest in a heterologous system. For that, pET plasmids carrying *ADH1* genes from *St. bombicola* and *W. cacticola* were cloned in an *Escherichia coli Rosetta* strain. The *E. coli Rosetta* (empty) contains an empty vector and is used as a control in this experiment since it does not have any *ADH1* gene.

To confirm if the genes were being overexpressed, *E. coli Rosetta* (empty), *E. coli Rosetta* pET19b*:ADH1* (*St. bombicola*) and *E. coli Rosetta* pET28a*:ADH1* (*W. cacticola*) were grown in 400 mL of LB medium and transcription was induced with 1 mM of IPTG for 4 h at 37°C. The overexpression of Adh1 proteins is shown on the SDS-PAGE gel of the Figure 3.9.

The expected size of Adh1 recombinant proteins is ~38 kDa. As it is possible to observe by the presence of a strong band at ~38 kDa, both *E. coli Rosetta* pET19b*:ADH1* (*St. bombicola*) (Figure 3.9 – B) and *E. coli Rosetta* pET28a*:ADH1 (W. cacticola)* (Figure 3.9 – C) strains are overproducing Adh1 when induced with 1 mM of IPTG. The absence of overexpression in the control culture, *E. coli Rosetta* (empty) (Figure 3.9 – A), also confirms that this intense band at ~38 kDa corresponds to Adh1.

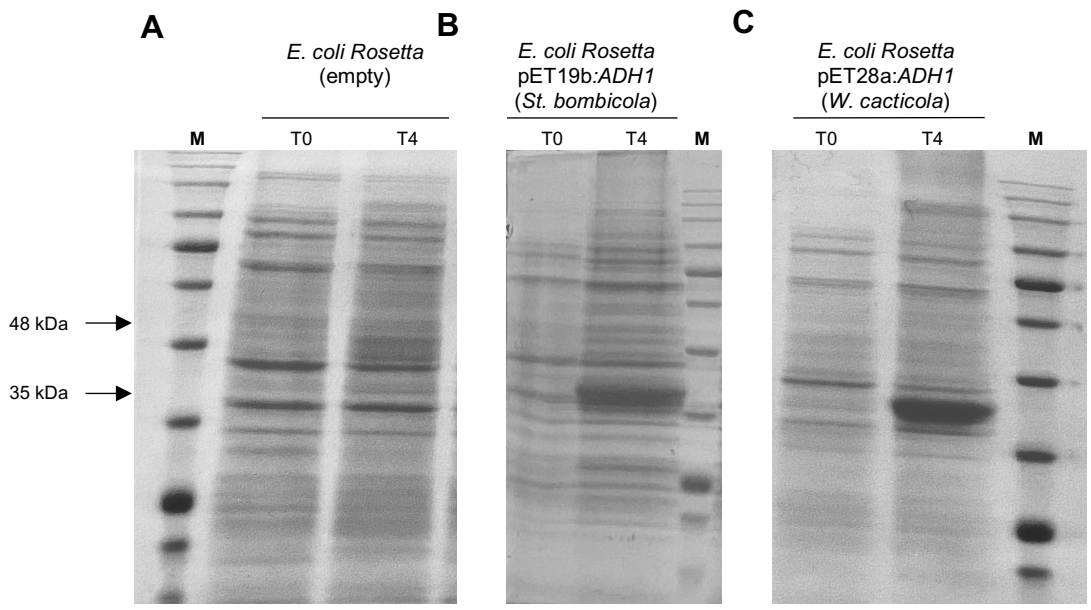


Figure 3.9. SDS-PAGE showing Adh1 overexpression after induction with 1 mM of IPTG for 4 h at 37°C. (A) *E. coli Rosetta* (empty). (B) *E. coli Rosetta* pET19b*:ADH1* (*St. bombicola*). (C) *E. coli Rosetta* pET28a*:ADH1 (W. cacticola)*. The overexpression of Adh1 is observable by a strong band at ~38 kDa on both *ADH1* molecular constructs. In *E. coli Rosetta* (empty) (control) there is no overexpression. M: Protein Marker (NZYColour Protein Marker II). T0: Before Induction with IPTG. T4: 4 h after induction with IPTG.

3.7.2. Solubilization of Adh1 proteins

3.7.2.1. Optimization of protein solubilization

It has been reported that the overexpression of heterologous proteins in *E. coli* often leads to their accumulation as insoluble Inclusion Bodies (IB). IB are composed by degraded and misfolded overproduced proteins (Baneyx, 1999). The aggregates are often formed since these proteins are being expressed at a high rate. This means less chaperones are available in the cell to assist correct folding, which promotes their sedimentation (Utekal et al., 2014).

There are different approaches to isolate and solubilize proteins from IB (Tsumoto et al., 2003). Nonetheless, the expression of a soluble protein is always more effective. This is because the solubilization process has to guarantee that the protein refolds into its correct conformation, which is sometimes difficult to achieve. To better increase the chances for the overexpression of a soluble protein, it is necessary to slow down the expression rate of the plasmid.

Utekal et al., 2014 were able to produce the first soluble Adh1 from *S. cerevisiae*, using *E. coli Rosetta* as the host organism. They were able to do so by decreasing the IPTG concentration to 0,5 mM (and reported IB formation while using higher concentrations). Besides this, the authors also decreased the induction temperature to 20°C and prolonged the induction time to 20 h.

The decrease of IPTG concentration and induction temperature leads to a slower expression rate of the protein of interest. The less IPTG there is to bind to the repressor, the less active the inducible system is. The temperature decrease is also important since it generally slows down the transcription and translation rates of the organism. It also weakens the strength of hydrophobic interactions that contribute to the sedimentation of the proteins (Baneyx & Mujacic, 2004). It is important to refer that in all of these experiments, the temperature was only decreased during induction. Before adding IPTG, the growth of *E. coli* cultures occurred at the optimal temperature of 37°C. Once the expression rate is slower, a lengthier induction time is necessary. Since less protein is being produced at a time, it is still crucial to guarantee that in the end of the induction time, there is enough protein for purification. To obtain an overexpressed and soluble Adh1 protein, different conditions were tested on a lower scale (using 50 mL LB medium instead of the 400 mL). These conditions varied on the IPTG concentrations, induction time and induction temperature.

3.7.2.2. The Adh1c from *W. cacticola* is possibly insoluble

Regarding the Adh1c from *W. cacticola*, it was observed that all of the tested conditions: IPTG concentrations (0,5 mM, 0,005 mM); induction time (4 h, 8 h, 20 h) and induction temperature (37°C, 28°C and 20°C) resulted in an insoluble protein. There are still different tests that can be made in order to try to solubilize the Adh1c protein of *W. cacticola*. A different approach is to let the *E. coli* cultures grow for a longer period of time before induction and use a shorter induction time. In this manner, there will be more cells to express the protein of interest. By harvesting these higher density cultures after a shorter period of time (for example 2 h), it is possible that it helps prevent the accumulation of the protein in IB. Auto-induction protocols can also be a possibility for obtaining soluble proteins (Grabski et al., 2005). Solubilization protocols can also be further tested. However, as it was mentioned, it is always easier to obtain a soluble protein by changing the upstream conditions of the experiment (IPTG concentrations, induction temperatures and induction time) than the downstream solubilization.

3.7.2.3. The overexpression of a soluble Adh1a from *St. bombicola*

Concerning the Adh1a from *St. bombicola*, only 0,5 mM of IPTG were used for the induction. The firstly tested conditions were: Induction for 4 h at 37°C, Induction for 8 h at 25°C, Induction for 8 h at 28°C, Induction for 20 h at 25°C. It was possible to observe an overexpressed Adh1a protein in all the cases. For this reason, all cultures were lysed by sonication. The resultant total extracts and respective supernatants were analysed by the SDS-PAGE that is presented in Figure 3.10.

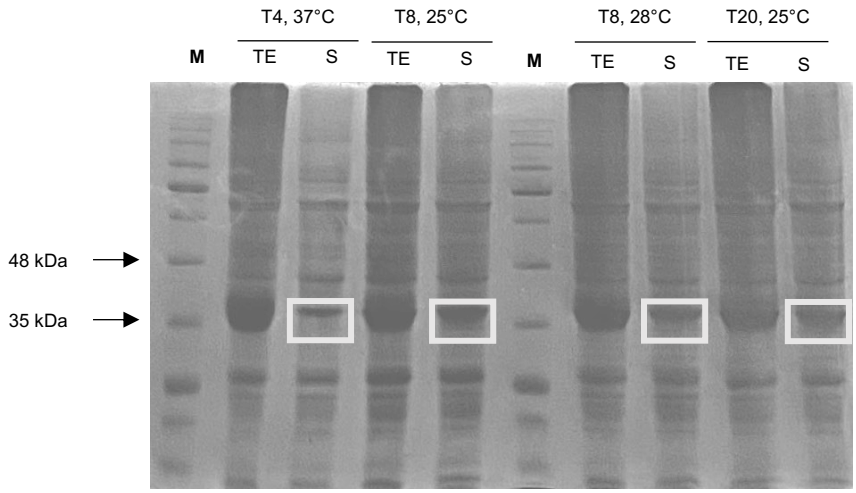


Figure 3.10. SDS-PAGE of the overexpression of a soluble Adh1a protein from *St. bombicola*. The induction of the expression of Adh1a proteins was performed by using 0,5 mM of IPTG. After cell lysis, the total extract (TE) and the respective supernatant (S) were analysed by SDS-PAGE. All tested conditions resulted in a soluble protein as it is possible to observe an overexpression band in all supernatants (grey rectangles). T4: 4 h after induction. T8: 8 h after induction. T20: 20 h after induction. M: Protein Marker (NZYColour Protein Marker II).

As it is possible to observe in Figure 3.10, all the analysed supernatants have a strong band at 38 kDa. This means the protein is soluble in all of the four conditions. Therefore, it was necessary to choose one of the conditions to perform the experiment on a larger scale (using 400 mL of LB medium) for further protein purification. After only 4 h of induction, the concentration of the overexpressed protein is clearly lower than the remaining three conditions. For this reason, this condition was excluded. Even though there is a strong band for the three conditions, the chosen condition was the 8 h of induction at 28°C. In this condition, it was clear that the ~38 kDa band is due to the overexpression of Adh1a and not the result of the overlapping of two bands.

The induction of the expression of Adh1a from *St. bombicola* was performed, for 8 h, with 0,5 mM IPTG at 28°C, in a large scale (using 400 mL of LB medium). The overexpression and presence of the overexpressed protein on the supernatant was confirmed by SDS-PAGE. Following this, the protein was purified. Eight aliquots of the elution buffer, containing 250 mM of Imidazole were retrieved but no band was visible at ~38 kDa by SDS-PAGE. To understand if there were any vestiges of Adh1a in the elution buffer, the enzymatic activities using NADH were performed on the eight portions, but no activity was detected. Given these results, a SDS-PAGE of the flow-through (FT) and the washing portions (A1, A2, A3) was performed and is shown in Figure 3.11.

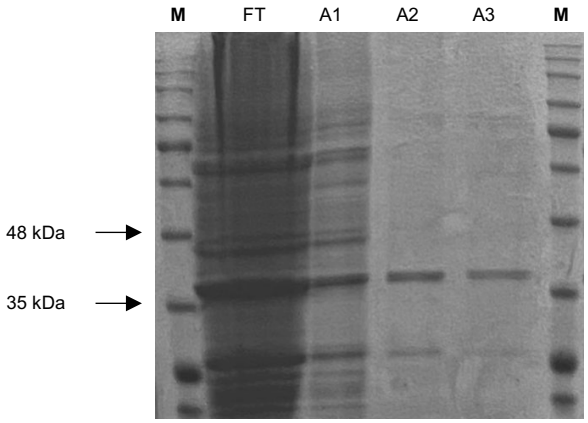


Figure 3.11. SDS-PAGE of the purification of the overexpressed Adh1a from *St. bombicola*. The overexpression of Adh1a was obtained by using 0,5 mM of IPTG for 8 h at 28°C. The soluble phase, after cell lysis, was used for protein purification. FT: Flow-through. A1, A2, A3: Washing fractions of the column, 3 mL each. M: Protein Marker (NZYColour Protein Marker II). The recombinant Adh1a protein did not bind to the column since most of it is on the FT.

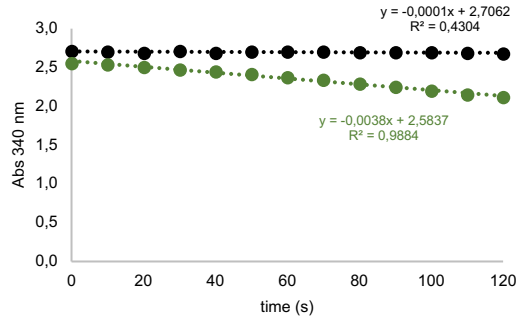
By observing the SDS-PAGE shown in Figure 3.11, it was noticeable that most of the Adh1a protein was eluted as soon as the column was charged (FT). The remaining washing fractions have gradually less protein. Concomitant with these results, FT showed a strong enzymatic activity for NADH. The remaining washing portions (A1, A2, A3) were also tested but did not have enough protein to perform the reaction.

This means that contrary to what as expected, the protein does not have a strong affinity for the column. In order to increase the affinity, the molecular construct must be redesigned and possibly be constructed in the opposite direction (for example, by placing the histidine tag in the N' terminal). Proteins often have a quaternary structure that could be blocking the histidine tag to adhere to the column.

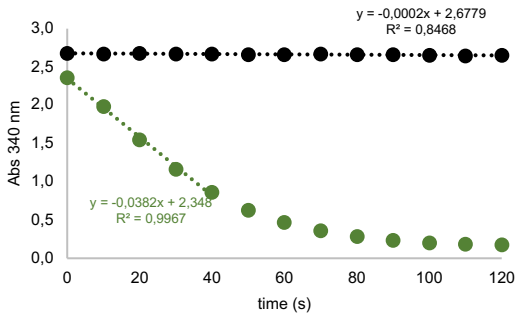
3.8. Enzymatic characterization of *E. coli* extracts

It was not possible to purify the Adh1 proteins of *St. bombicola* and *W. cacticola*. However, while the Adh1c from *W. cacticola* is possibly insoluble, the overexpression of a soluble Adh1a from *St. bombicola* was achieved. This means that it is possible to compare the enzymatic activities of the *E. coli Rosetta* pET19b*:ADH1* (*St. bombicola*) and *E. coli Rosetta* (empty) extracts, since the only difference between both extracts is the presence or absence of the Adh1a protein, respectively. The enzymatic assays were performed on the soluble phase of the *E. coli Rosetta* (empty) and on the *E. coli Rosetta* pET19b*:ADH1* (*St. bombicola*) extracts after sonication. The results regarding the direct reaction are represented in Figure 3.12. As for the inverse reaction it is represented in Figure 3.13.

A. NADH

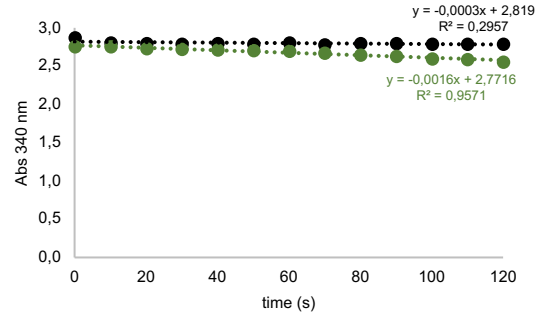


A1. *E. coli Rosetta* (empty)

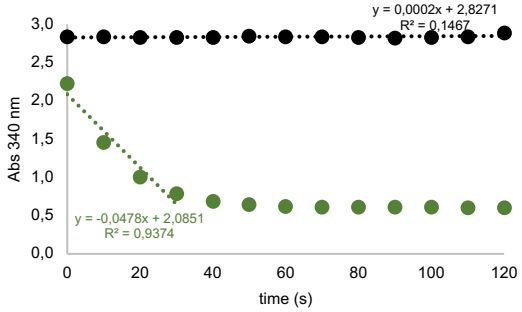


A2. *E. coli Rosetta* pET19b:*ADH1* (*St. bombicola*) **1:10**

B. NADPH



B1. *E. coli Rosetta* (empty)

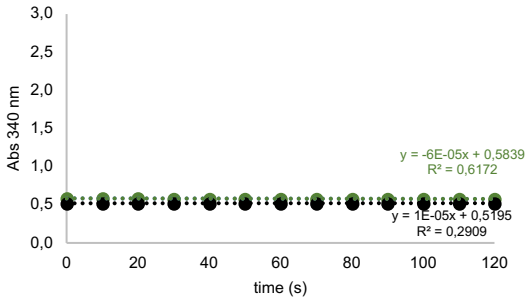


B2. *E. coli Rosetta* pET19b:*ADH1* (*St. bombicola*)

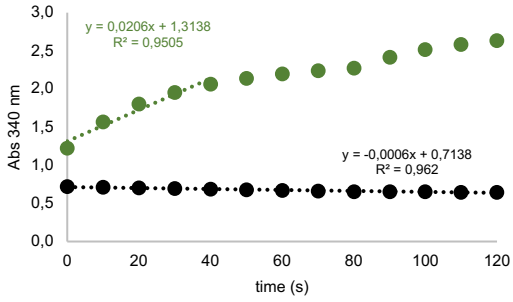
Figure 3.12. Alcohol dehydrogenase activity of *E. coli Rosetta* (empty) and *E. coli Rosetta* pET19b:*ADH1* (*St. bombicola*) protein extracts – direct reaction. (A) NADH and (B) NADPH consumption was measured at 340 nm for 120 sec (green dots). The blank of each reaction (before acetaldehyde addition) is represented by black dots. The *E. coli Rosetta* (empty) had a weak enzymatic activity when using NADH or NADPH as cofactors. The Adh1a of *St. bombicola* has stronger enzymatic activities when using NADH or NADPH as cofactors. With the exception of A2, in all enzymatic assays, 25 μL of the total extract were used. In the case of A2, a 1:10 dilution of the total extract was performed and 25 μL were used. This dilution was necessary in order to observe the reaction. For this reason, even though the module of the slope in B2 is higher, *St. bombicola* has a much stronger enzymatic activity while using NADH as a cofactor.

As it is possible to observe in Figure 3.12, the Adh1a enzyme is involved in the direct reaction of ethanol production from acetaldehyde. The *E. coli Rosetta* (empty) extract presents a weak enzymatic activity with NADH (Figure 3.12 – A1) and NADPH (Figure 3.12 – B1). This demonstrates that most of the measured activity in the *E. coli Rosetta* pET19b:*ADH1* (*St. bombicola*) extract is from Adh1a. The Adh1a from *St. bombicola* is able to use both NADH (Figure 3.12 – A2) and NADPH (Figure 3.12 – B2) as cofactors, having a preference for NADH. The same results were observed on previous studies of Gonçalves et al., 2018 while using *St. bombicola* total protein extracts (Gonçalves et al., 2018).

A. NAD$^+$

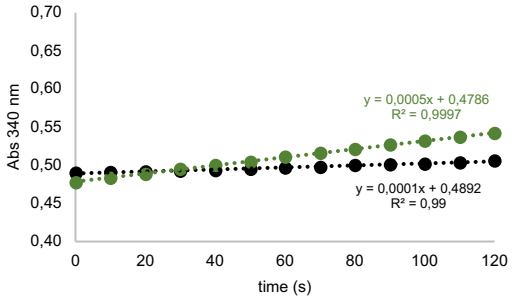


A1. *E. coli Rosetta* (empty)

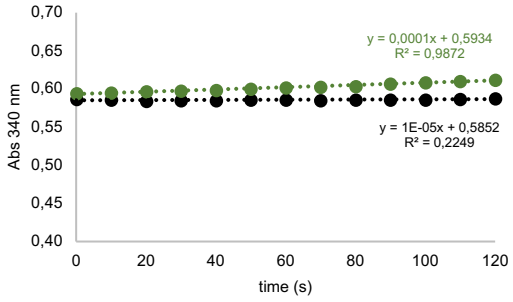


A2. *E. coli Rosetta* pET19b:*ADH1* (*St. bombicola*)

B. NADP$^+$



B1. *E. coli Rosetta* (empty)



B2. *E. coli Rosetta* pET19b:*ADH1* (*St. bombicola*)

Figure 3.13. Alcohol dehydrogenase activity of *E. coli Rosetta* (empty) and *E. coli Rosetta* pET19b*:ADH1* (*St. bombicola*) protein extracts – inverse reaction. (A) NADH and (B) NADPH production (by reduction of NAD$^+$ and NADP$^+$, respectively) was measured at 340 nm for 120 sec (green dots). The blank of each reaction (before ethanol addition) is represented by black dots. The *E. coli Rosetta* (empty) has a weak enzymatic activity whether by using NAD$^+$ or NADP$^+$ as a cofactor. The Adh1a of *St. bombicola* has a strong enzymatic activity while using NAD$^+$ as a cofactor. The comparison between B1 and B2 allows the understanding that Adh1a does not use this cofactor. In all enzymatic assays, 25 μL of the total extract were used.

As it is possible to observe in Figure 3.13, while using the NADP$^+$ as cofactor, both the *E. coli Rosetta* (empty) (Figure 3.13 – B1) and the *E. coli Rosetta* pET19b:*ADH1* (*St. bombicola*) extracts (Figure 3.13 – B2) present a very weak enzymatic activity. Since there is a residual activity detected in both extracts and slightly stronger for the *E. coli Rosetta* (empty) extract, this indicates that the Adh1a protein has negligent or no activity in the reverse reaction using NADP$^+$. This residual activity could be the result of unspecific reactions from *E. coli* proteins that are also present on the extract. This represents an example of what has been aforementioned regarding the difficulties of enzymatic characterization in protein extracts. If the comparison was not possible, the detected NADP$^+$ activity in *E. coli Rosetta* pET19b:*ADH1* (*St. bombicola*) extract (Figure 3.13 – B2) would be wrongly attributed to Adh proteins. In this case, having the two extracts for comparison is essential.

The *E. coli Rosetta* (empty) extract does not have any measurable enzymatic activity with NAD$^+$ (Figure 3.13 – A1). This strongly indicates that the detected enzymatic activity, using the same cofactor in the *E. coli Rosetta* pET19b:*ADH1* (*St. bombicola*) extract (Figure 3.13 – A2) is from Adh1a. This confirms that besides performing the direct reaction, the Adh1a enzyme is also involved in the inverse reaction,

converting ethanol into acetaldehyde. These results are in line with what was previously observed in the *St. bombicola* mutant that had its *ADH1* gene deleted. As a consequence of the *ADH1* deletion, *St. bombicola* lost its ability to assimilate ethanol (Gonçalves et al., 2018).

In short, the Adh1a protein from *St. bombicola* is the main enzyme involved in the interconversion of acetaldehyde and ethanol, in contrast to what has been observed for *S. cerevisiae*, whose Adh1 is predominantly involved in the direct reaction. Besides this, the Adh1a protein from *St. bombicola* uses both NADH and NADPH as cofactors. This is also a different characteristic from the *S. cerevisiae* Adh1 protein that only uses NADH (de Smidt et al., 2008, 2011). The comparison between extracts, having *E. coli Rosetta* (empty) as control, provide an important resource for the characterization of Adh proteins. This experiment can be used as an alternative for Adh1 characterization when overexpression and solubility is achieved, but purification is not successful.

4. Conclusion and future perspectives

The rapid increase of available genomic data, along with the development of comparative genomics tools, have allowed the uncovering of more and more HGT events (Husnik & McCutcheon, 2017). The W/S clade is known to have a higher than usual incidence of HGT-derived genes (Gonçalves et al., 2018; Shen et al., 2018). It is clear that this mechanism is an important driving force for genome evolution in the W/S clade and that its overall importance for eukaryotic evolution may have been underestimated (Fitzpatrick, 2011). Given the current *in silico* data, the observed ethanol metabolism profiles, and the enzymatic assays, it is possible to infer that the interdomain HGT of bacterial *ADH* genes have been successful for some species of the W/S clade. This means that the bacterial genes became established and functional in eukaryotic settings, which was mainly observed on species that produced large amounts of ethanol.

The ability to produce ethanol by the W/S-clade yeasts is in an important advantage for biotechnological purposes, for example, in the production of alcoholic beverages. Traditionally, the mostly used yeast is *Saccharomyces cerevisiae*. However, non-conventional yeasts as the members of the W/S clade, can also provide several advantages. These species, not only produce less ethanol but also a range of secondary metabolites, such as glycerol. These represent characteristics that are often found desirable for the production of alcoholic drinks (Binati et al., 2020; Englezos et al., 2016; Gonçalves et al., 2019; Masneuf-Pomarede et al., 2016).

Besides this, W/S-clade yeasts are also fructophilic. Since *S. cerevisiae* is glucophilic, which means it has an overall preference for glucose as a carbon source, there is an advantage of using both yeasts for mixed fermentations. By performing these, the spoilage of fructose in stuck fermentations is avoided and therefore economical losses are decreased. Currently, one of the most promising species for the mixed fermentations with *S. cerevisiae* is a W/S-clade yeast, *Starmerella bacillaris* (Binati et al., 2020; Englezos et al., 2016; Masneuf-Pomarede et al., 2016).

Starmerella bacillaris belongs to the subgroup A (Gonçalves et al., 2018). Interestingly, all the species in which the re-establishment of the alcoholic fermentation pathway was confirmed experimentally, also belong to this subgroup and have horizontally acquired the *ADH1a* type. The role of *ADH1b* and *ADH1c* in alcoholic fermentation remains obscure. However, most of the evaluated species that harbour these genes, could at least assimilate ethanol. It is still not clear whether this capacity was restored by the acquisition of *ADH* genes, or if it is the result of other unspecific dehydrogenases. For this reason, the purification and characterization of Adh1 proteins is crucial.

Regarding Aro10, all the results point to the possibility that this enzyme has evolved to replace Pdc1 in the conversion of pyruvate to acetaldehyde in species that regained the ability to produce ethanol. While Aro10 is present across all the evaluated species of the W/S clade (except from *Wickerhamiella*

slavikovae), the native Pdc1 has been completely lost. However, the substitutional role of Aro10 has only been biochemically verified in *Starmerella bombicola*, whose ethanol-producing capacity was completely obliterated by the elimination of this gene (Gonçalves et al., 2018). It is very likely the same is occurring in other W/S-clade species. However, to further confirm this, it is necessary to study Aro10 deletion mutants of these species. So far, *St. bombicola* (Gonçalves et al., 2018) and *W. sorbophila* (Lee et al., 2018) are the only W/S-clade species to have been successfully genetically engineered.

The W/S clade is, indeed, a very diverse lineage, especially the *Wickerhamiella* genus, given the number of independent HGT events that have occurred. It is highly probable that as more W/S-clade genomes become available, more bacterial *ADH1* and *ADH6* genes, and more species lacking *PDC1* but harbouring *ARO10*, will be found. For further studies, it would also be very interesting to find possible pseudogenes of the native yeast *ADH1*, *ADH6* and *PDC1* on the W/S-clade genomes. This would help to better understand the mechanisms of gene loss and when these took place. Overall, the W/S clade represents a promising model lineage for the study of HGT as an important driver of genetic diversity and evolution among eukaryotes. Even though this mechanism has been widely observed and described in prokaryotes, little is still known about how it operates in eukaryotes.

References

1. Albalat, R., & Cañestro, C. (2016). Evolution by gene loss. *Nature Reviews Genetics*, *17*(7), 379-391. doi: 10.1038/nrg.2016.39

2. Baneyx, F. (1999). Recombinant protein expression in *Escherichia coli*. *Current Opinion In Biotechnology*, *10*(5), 411-421. doi: 10.1016/s0958-1669(99)00003-8

3. Baneyx, F., & Mujacic, M. (2004). Recombinant protein folding and misfolding in *Escherichia coli*. *Nature Biotechnology*, *22*(11), 1399-1408. doi: 10.1038/nbt1029

4. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., & Kulikov, A. S. et al. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal Of Computational Biology*, *19*(5), 455-477. doi: 10.1089/cmb.2012.0021

5. Binati, R. L., Lemos Junior, W. J. F., Luzzini, G., Slaghenaufi, D., Ugliano, M., & Torriani, S. (2020). Contribution of non-*Saccharomyces* yeasts to wine volatile and sensory diversity: A study on *Lachancea thermotolerans*, *Metschnikowia* spp. and *Starmerella bacillaris* strains isolated in Italy. *International Journal Of Food Microbiology*, *318*, 108470. doi: 10.1016/j.ijfoodmicro.2019.108470

6. Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114-2120. doi: 10.1093/bioinformatics/btu170

7. Bornhorst, J. A., & Falke, J. J. (2000). [16] Purification of proteins using polyhistidine affinity tags. *Methods In Enzymology*, *326*, 245-254, doi: 10.1016/s0076-6879(00)26058-8

8. Cabral, S., Prista, C., Loureiro-Dias, M. C., & Leandro, M. J. (2015). Occurrence of *FFZ* genes in yeasts and correlation with fructophilic behaviour. *Microbiology*, *161*(10), 2008-2018. doi: 10.1099/mic.0.000154

9. Capella-Gutierrez, S., Silla-Martinez, J. M., & Gabaldon, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, *25*(15), 1972-1973. doi: 10.1093/bioinformatics/btp348

10. Chikhi, R., & Medvedev, P. (2013). Informed and automated k-mer size selection for genome assembly. *Bioinformatics*, *30*(1), 31-37. doi: 10.1093/bioinformatics/btt310

11. Daniel, H. -M., & Meyer, W. (2003). Evaluation of ribosomal RNA and actin gene sequences for the identification of ascomycetous yeasts. *International Journal Of Food Microbiology*, *86*(1-2), 61-78. doi: 10.1016/s0168-1605(03)00248-4

12. Dashko, S., Zhou, N., Compagno, C., & Piškur, J. (2014). Why, when, and how did yeast evolve alcoholic fermentation?. *FEMS Yeast Research*, *14*(6), 826-832. doi: 10.1111/1567-1364.12161

13. De Deken, R. H. (1966). The Crabtree Effect: A Regulatory System in Yeast. *Journal Of General Microbiology*, *44*(2), 149-156. doi: 10.1099/00221287-44-2-149

14. de Smidt, O., du Preez, J. C., & Albertyn, J. (2008). The alcohol dehydrogenases of *Saccharomyces cerevisiae*: a comprehensive review. *FEMS Yeast Research*, *8*(7), 967-978. doi: 10.1111/j.1567-1364.2008.00387.x

15. de Smidt, O., du Preez, J. C., & Albertyn, J. (2011). Molecular and physiological aspects of alcohol dehydrogenases in the ethanol metabolism of *Saccharomyces cerevisiae*. *FEMS Yeast Research*, *12*(1), 33-47. doi: 10.1111/j.1567-1364.2011.00760.x

16. de Vega, C., Albaladejo, R. G., Guzmán, B., Steenhuisen, S. -.L, Johnson, S. D., Herrera, C. M, & Lachance, M. -A. (2017). Flowers as a reservoir of yeast diversity: description of *Wickerhamiella nectarea* f.a. sp. nov., and *Wickerhamiella natalensis* f.a. sp. nov. from South African flowers and pollinators, and transfer of related *Candida* species to the genus *Wickerhamiella* as new combinations. *FEMS Yeast Research*, *17*(5). doi: 10.1093/femsyr/fox054

17. Diamantino, R. M. F. (2020). *Caracterização bioquímica de álcool desidrogenase de origem bacteriana em leveduras não convencionais.* [Unpublished Graduation thesis]. Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Caparica, Portugal

18. Dickinson, J. R., Salgado, L. E. J., & Hewlins, M. J. E. (2003). The Catabolism of Amino Acids to Long Chain and Complex Alcohols in *Saccharomyces cerevisiae*. *Journal Of Biological Chemistry*, *278*(10), 8028-8034. doi: 10.1074/jbc.m211914200

19. Doolittle, W. F. (1998). You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends In Genetics*, *14*(8), 307-311. doi: 10.1016/s0168-9525(98)01494-2

20. Drewke, C., Thielen, J., & Ciriacy, M. (1990). Ethanol formation in *adh0* mutants reveals the existence of a novel acetaldehyde-reducing activity in *Saccharomyces cerevisiae*. *Journal Of Bacteriology*, *172*(7), 3909-3917. doi: 10.1128/jb.172.7.3909-3917.1990

21. Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*(1), 238. doi: 10.1186/s13059-019-1832-y

22. Endo, A., & Okada, S. (2008). Reclassification of the genus *Leuconostoc* and proposals of *Fructobacillus fructosus* gen. nov., comb. nov., *Fructobacillus durionis* comb. nov., *Fructobacillus ficulneus* comb. nov. and *Fructobacillus pseudoficulneus* comb. nov. *International Journal Of Systematic And Evolutionary Microbiology*, *58*(Pt 9), 2195-2205. doi: 10.1099/ijs.0.65609-0

23. Endo, A., & Salminen, S. (2013). Honeybees and beehives are rich sources for fructophilic lactic acid bacteria. *Systematic And Applied Microbiology*, *36*(6), 444-448. doi: 10.1016/j.syapm.2013.06.002

24. Endo, A., Futagawa-Endo, Y., & Dicks, L. M. T. (2009). Isolation and characterization of fructophilic lactic acid bacteria from fructose-rich niches. *Systematic And Applied Microbiology*, *32*(8), 593-600. doi: 10.1016/j.syapm.2009.08.002

25. Endo, A., Irisawa, T., Futagawa-Endo, Y., Takano, K., du Toit, M., Okada, S., & Dicks, L. M. T. (2012). Characterization and emended description of *Lactobacillus kunkeei* as a fructophilic lactic acid bacterium. *International Journal Of Systematic And Evolutionary Microbiology*, *62*(Pt_3), 500-504. doi: 10.1099/ijs.0.031054-0

26. Endo, A., Maeno, S., Tanizawa, Y., Kneifel, W., Arita, M., Dicks, L., & Salminen, S. (2018). Fructophilic Lactic Acid Bacteria, a Unique Group of Fructose-Fermenting Microbes. *Applied And Environmental Microbiology*, *84*(19), e01290-18. doi: 10.1128/aem.01290-18

27. Endo, A., Tanaka, N., Oikawa, Y., Okada, S., & Dicks, L. (2013). Fructophilic Characteristics of *Fructobacillus* spp. may be due to the Absence of an Alcohol/Acetaldehyde Dehydrogenase Gene (*adhE*). *Current Microbiology*, *68*(4), 531-535. doi: 10.1007/s00284-013-0506-3

28. Englezos, V., Rantsiou, K., Cravero, F., Torchio, F., Ortiz-Julien, A., & Gerbi, V. et al. (2016). *Starmerella bacillaris* and *Saccharomyces cerevisiae* mixed fermentations to reduce ethanol content in wine. *Applied Microbiology and Biotechnology*, *100*(12), 5515-5526. doi: 10.1007/s00253-016-7413-z

29. Feng, B., Lin, Y., Zhou, L., Guo, Y., Friedman, R., & Xia, R. et al. (2017). Reconstructing Yeasts Phylogenies and Ancestors from Whole Genome Data. *Scientific Reports*, *7*(1), 15209. doi: 10.1038/s41598-017-15484-5

30. Fitzpatrick, D. A. (2011). Horizontal gene transfer in fungi. *FEMS Microbiology Letters*, *329*(1), 1-8. doi: 10.1111/j.1574-6968.2011.02465.x

31. Galagan, J. E., Henn, M. R., Ma, L. -J., Cuomo, C. A., & Birren, B. (2005). Genomics of the fungal kingdom: Insights into eukaryotic biology. *Genome Research*, *15*(12), 1620-1631. doi: 10.1101/gr.3767105

32. Garvão, F. (2020). *Caracterização bioquímica de álcool desidrogenases de origem bacteriana em leveduras não convencionais*. [Unpublished Graduation thesis]. Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Caparica, Portugal

33. Geiler-Samerotte, K. A., Bauer, C. R., Li, S., Ziv, N., Gresham, D., & Siegal, M. L. (2013). The details in the distributions: why and how to study phenotypic variability. *Current Opinion In Biotechnology*, *24*(4), 752-759. doi: 10.1016/j.copbio.2013.03.010

34. Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., & Feldmann, H. et al. (1996). Life with 6000 Genes. *Science*, *274*(5287), 546-567. doi: 10.1126/science.274.5287.546

35. Gonçalves, C., & Gonçalves, P. (2019). Multilayered horizontal operon transfers from bacteria reconstruct a thiamine salvage pathway in yeasts. *Proceedings Of The National Academy Of Sciences*, *116*(44), 22219-22228. doi: 10.1073/pnas.1909844116

36. Gonçalves, C., Coelho, M. A., Salema-Oom, M., & Gonçalves, P. (2015). Stepwise Functional Evolution in a Fungal Sugar Transporter Family. *Molecular Biology And Evolution*, *33*(2), 352-366. doi: 10.1093/molbev/msv220

37. Gonçalves, C., Ferreira, C., Gonçalves, L. G., Turner, D. L., Leandro, M. J., & Salema-Oom, M. et al. (2019). A New Pathway for Mannitol Metabolism in Yeasts Suggests a Link to the Evolution of Alcoholic Fermentation. *Frontiers In Microbiology*, *10*(2510). doi: 10.3389/fmicb.2019.02510

38. Gonçalves, C., Wisecaver, J. H., Kominek, J., Oom, M. S., Leandro, M. J., & Shen, X. -.X et al. (2018). Evidence for loss and reacquisition of alcoholic fermentation in a fructophilic yeast lineage. *Elife*, *7*, e33034. doi: 10.7554/elife.33034

39. Gonçalves, P., Gonçalves, C., Brito, P. H., & Sampaio, J. P. (2020). The *Wickerhamiella*/*Starmerella* clade-A treasure trove for the study of the evolution of yeast metabolism. *Yeast*, *37*(4), 313-320. doi: 10.1002/yea.3463

40. Gordon, J. L., Byrne, K. P., & Wolfe, K. H. (2009). Additions, Losses, and Rearrangements on the Evolutionary Route from a Reconstructed Ancestor to the Modern *Saccharomyces cerevisiae* Genome. *Plos Genetics*, *5*(5), e1000485. doi: 10.1371/journal.pgen.1000485

41. Grabski, A., Mehler, M., & Drott, D. (2005). The Overnight Express Autoinduction System: High-density cell growth and protein expression while you sleep. *Nature Methods*, *2*(3), 233-235. doi: 10.1038/nmeth0305-233

42. Hittinger, C. T., Rokas, A., Bai, F. -Y., Boekhout, T., Gonçalves, P., & Jeffries, T. W. et al. (2015). Genomics and the making of yeast biodiversity. *Current Opinion In Genetics & Development*, *35*, 100-109. doi: 10.1016/j.gde.2015.10.008

43. Hohmann, S. (1991). Characterization of *PDC6*, a third structural gene for pyruvate decarboxylase in Saccharomyces cerevisiae. *Journal Of Bacteriology*, *173*(24), 7963-7969. doi: 10.1128/jb.173.24.7963-7969.1991

44. Hohmann, S., & Cederberg, H. (1990). Autoregulation may control the expression of yeast pyruvate decarboxylase structural genes *PDC1* and *PDC5*. *European Journal Of Biochemistry*, *188*(3), 615-621. doi: 10.1111/j.1432-1033.1990.tb15442.x

45. Huang, J. (2013). Horizontal gene transfer in eukaryotes: The weak-link model. *Bioessays*, *35*(10), 868-875. doi: 10.1002/bies.201300007

46. Husnik, F., & McCutcheon, J. P. (2017). Functional horizontal gene transfer from bacteria to eukaryotes. *Nature Reviews Microbiology*, *16*(2), 67-79. doi: 10.1038/nrmicro.2017.137

47. Ida, Y., Furusawa, C., Hirasawa, T., & Shimizu, H. (2012). Stable disruption of ethanol production by deletion of the genes encoding alcohol dehydrogenase isozymes in *Saccharomyces cerevisiae*. *Journal Of Bioscience And Bioengineering*, *113*(2), 192-195. doi: 10.1016/j.jbiosc.2011.09.019

48. Iino, T., Suzuki, R., Kosako, Y., Ohkuma, M., Komagata, K., & Uchimura, T. (2012). *Acetobacter okinawensis* sp. nov., *Acetobacter papayae* sp. nov., and *Acetobacter persicus* sp. nov.; novel acetic acid bacteria isolated from stems of sugarcane, fruits, and a flower in Japan. *The Journal Of General And Applied Microbiology*, *58*(3), 235-243. doi: 10.2323/jgam.58.235

49. Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology And Evolution*, *30*(4), 772-780. doi: 10.1093/molbev/mst010

50. Keeling, P. J., & Palmer, J. D. (2008). Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics*, *9*(8), 605-618. doi: 10.1038/nrg2386

51. Kneen, M. M., Stan, R., Yep, A., Tyler, R. P., Saehuan, C., & McLeish, M. J. (2011). Characterization of a thiamin diphosphate-dependent phenylpyruvate decarboxylase from *Saccharomyces cerevisiae*. *FEBS Journal*, *278*(11), 1842-1853. doi: 10.1111/j.1742-4658.2011.08103.

52. Kominek, J., Doering, D. T., Opulente, D. A., Shen, X. -X., Zhou, X., & DeVirgilio, J. et al. (2019). Eukaryotic Acquisition of a Bacterial Operon. *Cell*, *176*(6), 1356-1366.e10. doi: 10.1016/j.cell.2019.01.034

53. Lachance, M. -A. (2006) Yeast Biodiversity: How Many and How Much?. In: Péter G., Rosa C. (eds) Biodiversity and Ecophysiology of Yeasts. *The Yeast Handbook*. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-30985-3_1

54. Lachance, M. -A., Starmer, W. T., Rosa, C. A., Bowles, J. M., Barker, J. S. F., & Janzen, D. H. (2001). Biogeography of the yeasts of ephemeral flowers and their insects. *FEMS Yeast Research*, *1*(1), 1-8. doi: 10.1111/j.1567-1364.2001.tb00007.x

55. Larroy, C., Fernández, M. R., González, E., Parés, X., & Biosca, J. A. (2002a). Characterization of the *Saccharomyces cerevisiae* YMR318C (*ADH6*) gene product as a broad specificity NADPH-dependent alcohol dehydrogenase: relevance in aldehyde reduction. *Biochemical Journal*, *361*(Pt 1), 163-172. doi: 10.1042/0264-6021:3610163

56. Larroy, C., Parés, X., & Biosca, J. A. (2002b). Characterization of a *Saccharomyces cerevisiae* NADP(H)-dependent alcohol dehydrogenase (ADHVII), a member of the cinnamyl alcohol dehydrogenase family. *European Journal Of Biochemistry*, *269*(22), 5738-5745. doi: 10.1046/j.1432-1033.2002.03296.x

57. Leandro, M. J., Cabral, S., Prista, C., Loureiro-Dias, M. C., & Sychrová, H. (2014). The High-Capacity Specific Fructose Facilitator ZrFfz1 Is Essential for the Fructophilic Behavior of *Zygosaccharomyces rouxii* CBS 732T. *Eukaryotic Cell*, *13*(11), 1371-1379. doi: 10.1128/ec.00137-14

58. Lee, H., Han, C., Lee, H. -W., Park, G., Jeon, W., Ahn, J., & Lee, H. (2018). Development of a promising microbial platform for the production of dicarboxylic acids from biorenewable resources. *Biotechnology For Biofuels*, *11*(1). doi: 10.1186/s13068-018-1310-x

59. Letunic, I., & Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research*, gkab301, doi: 10.1093/nar/gkab301

60. Li, W., & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), 1658-1659. doi: 10.1093/bioinformatics/btl158

61. Lindsey, A. R. I., & Newton, I. L. G. (2019). Some Like it HOT: Horizontal Operon Transfer. *Cell*, *176*(6), 1243-1245. doi: 10.1016/j.cell.2019.02.007

62. Maeno, S., Kajikawa, A., Dicks, L., & Endo, A. (2019). Introduction of bifunctional alcohol/acetaldehyde dehydrogenase gene (*adhE*) in *Fructobacillus fructosus* settled its fructophilic characteristics. *Research In Microbiology*, *170*(1), 35-42. doi: 10.1016/j.resmic.2018.09.004

63. Maeno, S., Tanizawa, Y., Kanesaki, Y., Kubota, E., Kumar, H., & Dicks, L. et al. (2016). Genomic characterization of a fructophilic bee symbiont *Lactobacillus kunkeei* reveals its niche-specific adaptation. *Systematic and Applied Microbiology*, *39*(8), 516-526. doi: 10.1016/j.syapm.2016.09.006

64. Marcet-Houben, M., & Gabaldón, T. (2010). Acquisition of prokaryotic genes by fungal genomes. *Trends In Genetics*, *26*(1), 5-8. doi: 10.1016/j.tig.2009.11.007

65. Masneuf-Pomarede, I., Bely, M., Marullo, P., & Albertin, W. (2016). The Genetics of Non-conventional Wine Yeasts: Current Knowledge and Future Challenges. *Frontiers In Microbiology*, *6*, 1563. doi: 10.3389/fmicb.2015.01563

66. Masud, U., Matsushita, K., & Theeragool, G. (2011). Molecular cloning and characterization of two inducible NAD^+-adh genes encoding NAD^+-dependent alcohol dehydrogenases from *Acetobacter pasteurianus* SKU1108. *Journal of Bioscience And Bioengineering*, *112*(5), 422-431. doi: 10.1016/j.jbiosc.2011.07.020

67. Milner, D. S., Attah, V., Cook, E., Maguire, F., Savory, F. R., & Morrison, M. et al. (2019). Environment-dependent fitness gains can be driven by horizontal gene transfer of transporter-encoding genes. *Proceedings of The National Academy Of Sciences*, *116*(12), 5613-5622. doi: 10.1073/pnas.1815994116

68. Minh, B. Q., Nguyen, M. A. T., & von Haeseler, A. (2013). Ultrafast Approximation for Phylogenetic Bootstrap. *Molecular Biology And Evolution*, *30*(5), 1188-1195. doi: 10.1093/molbev/mst024

69. Mitrophanov, A. Y., & Borodovsky, M. (2006). Statistical significance in biological sequence analysis. *Briefings In Bioinformatics*, *7*(1), 2-24. doi: 10.1093/bib/bbk001

70. Nguyen, L. -T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2014). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology And Evolution*, *32*(1), 268-274. doi: 10.1093/molbev/msu300

71. Pombert, J. -F., Selman, M., Burki, F., Bardell, F. T., Farinelli, L., & Solter, L. F. et al. (2012). Gain and loss of multiple functionally related, horizontally transferred genes in the reduced genomes of two microsporidian parasites. *Proceedings Of The National Academy of Sciences*, *109*(31), 12638-12643. doi: 10.1073/pnas.1205020109

72. Richards, T. A., & Talbot, N. J. (2013). Horizontal gene transfer in osmotrophs: playing with public goods. *Nature Reviews Microbiology*, *11*(10), 720-727. doi: 10.1038/nrmicro3108

73. Robbertse, B., Reeves, J. B., Schoch, C. L., & Spatafora, J. W. (2006). A phylogenomic analysis of the Ascomycota. *Fungal Genetics And Biology*, *43*(10), 715-725. doi: 10.1016/j.fgb.2006.05.001

74. Romagnoli, G., Luttik, M. A. H., Kötter, P., Pronk, J. T., & Daran, J. -M. (2012). Substrate Specificity of Thiamine Pyrophosphate-Dependent 2-Oxo-Acid Decarboxylases in Saccharomyces cerevisiae. *Applied And Environmental Microbiology*, *78*(21), 7538-7548. doi: 10.1128/aem.01675-12

75. Rosano, G. L., & Ceccarelli, E. A. (2014). Recombinant protein expression in *Escherichia coli*: advances and challenges. *Frontiers In Microbiology*, *5*, 172 doi: 10.3389/fmicb.2014.00172

76. Santos, A. R. O., Leon, M. P., Barros, K. O., Freitas, L. F. D., Hughes, A. F. S., & Morais, P. B. et al. (2018). *Starmerella camargoi* f.a., sp. nov., *Starmerella ilheusensis* f.a., sp. nov., *Starmerella*

litoralis f.a., sp. nov., *Starmerella opuntiae* f.a., sp. nov., *Starmerella roubikii* f.a., sp. nov. and *Starmerella vitae* f.a., sp. nov., isolated from flowers and bees, and transfer of related *Candida* species to the genus *Starmerella* as new combinations. *International Journal of Systematic And Evolutionary Microbiology*, *68*(4), 1333-1343. doi: 10.1099/ijsem.0.002675

77. Sevillya, G., Adato, O., & Snir, S. (2020). Detecting horizontal gene transfer: a probabilistic approach. *BMC Genomics*, *21*(S1). doi: 10.1186/s12864-019-6395-5

78. Shen, X. -X., Opulente, D. A., Kominek, J., Zhou, X., Steenwyk, J. L., & Buh, K. V. et al. (2018). Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum. *Cell*, *175*(6), 1533-1545.e20. doi: 10.1016/j.cell.2018.10.023

79. Sibbald, S. J., Eme, L., Archibald, J. M., & Roger, A. (2020). Lateral Gene Transfer Mechanisms and Pan-genomes in Eukaryotes. *Trends In Parasitology*, *36*(11), 927-941. doi: 10.1016/j.pt.2020.07.014

80. Soucy, S. M., Huang, J., & Gogarten, J. P. (2015). Horizontal gene transfer: building the web of life. *Nature Reviews Genetics*, *16*(8), 472-482. doi: 10.1038/nrg3962

81. Stanke, M., Diekhans, M., Baertsch, R., & Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*, *24*(5), 637-644. doi: 10.1093/bioinformatics/btn013

82. Tian, R. -M., Cai, L., Zhang, W. -P., Cao, H. -L., & Qian, P. -Y. (2015). Rare Events of Intragenus and Intraspecies Horizontal Transfer of the 16S rRNA Gene. *Genome Biology And Evolution*, *7*(8), 2310-2320. doi: 10.1093/gbe/evv143

83. Tsumoto, K., Ejima, D., Kumagai, I., & Arakawa, T. (2003). Practical considerations in refolding proteins from inclusion bodies. *Protein Expression And Purification*, *28*(1), 1-8. doi: 10.1016/s1046-5928(02)00641-1

84. Utekal, P., Tóth, C., Illésová, A., Koiš, P., Bocánová, L., & Turňa, J. et al. (2014). Expression of soluble *Saccharomyces cerevisiae* alcohol dehydrogenase in *Escherichia coli* applicable to oxido-reduction bioconversions. *Biologia*, *69*(6), 722-726. doi: 10.2478/s11756-014-0376-6

85. Will, J. L., Kim, H. S., Clarke, J., Painter, J. C., Fay, J. C., & Gasch, A. P. (2010). Incipient Balancing Selection through Adaptive Loss of Aquaporins in Natural *Saccharomyces cerevisiae* Populations. *Plos Genetics*, *6*(4), e1000893. doi: 10.1371/journal.pgen.1000893

86. Yang, Z., & Rannala, B. (2012). Molecular phylogenetics: principles and practice. *Nature Reviews Genetics*, *13*(5), 303-314. doi: 10.1038/nrg3186

87. Yvert, G., Ohnuki, S., Nogami, S., Imanaga, Y., Fehrmann, S., Schacherer, J., & Ohya, Y. (2013). Single-cell phenomics reveals intra-species variation of phenotypic noise in yeast. *BMC Systems Biology*, *7*(1), 54. doi: 10.1186/1752-0509-7-54

88. Zamora, F. (2009). Biochemistry of Alcoholic Fermentation. In: Moreno-Arribas M.V., Polo M.C. (eds.), *Wine Chemistry And Biochemistry*. Springer, New York, NY. https://doi.org/10.1007/978-0-387-74118-5_1

89. Zea, L., Serratosa, M. P., Mérida, J., & Moyano, L. (2015). Acetaldehyde as Key Compound for the Authenticity of Sherry Wines: A Study Covering 5 Decades. *Comprehensive Reviews in Food Science And Food Safety*, *14*(6), 681-693. doi: 10.1111/1541-4337.12159

Appendix

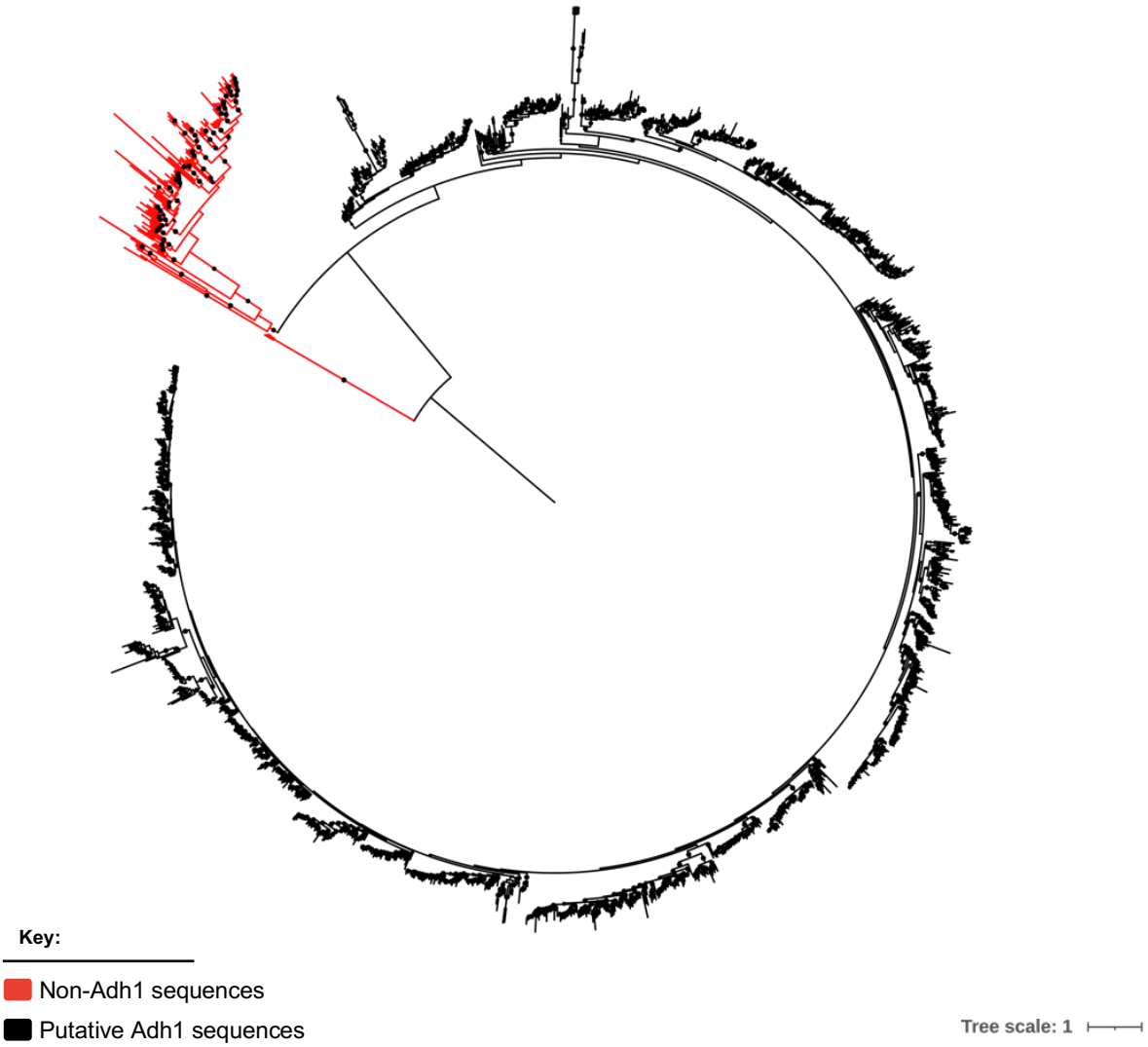


Figure A1. Preliminary Adh1 phylogeny. The preliminary tree of Adh1 was constructed with the top 5.000 BLASTp hits, using the *Starmerella bombicola* Adh1 sequence as query. The predicted Adh1 and Adh6 protein sequences from the W/S-clade genomes were added. Two distinct groups are observed in this phylogeny. Some protein sequences from the W/S-clade genomes, form the separate group of non-Adh1 sequences (red). This group does not cluster with the putative Adh1 sequences (black) and was removed in order to construct the final Adh1 phylogeny that is observed in Figure 3.2. Bootstrap values are indicated with black circles (>90%).

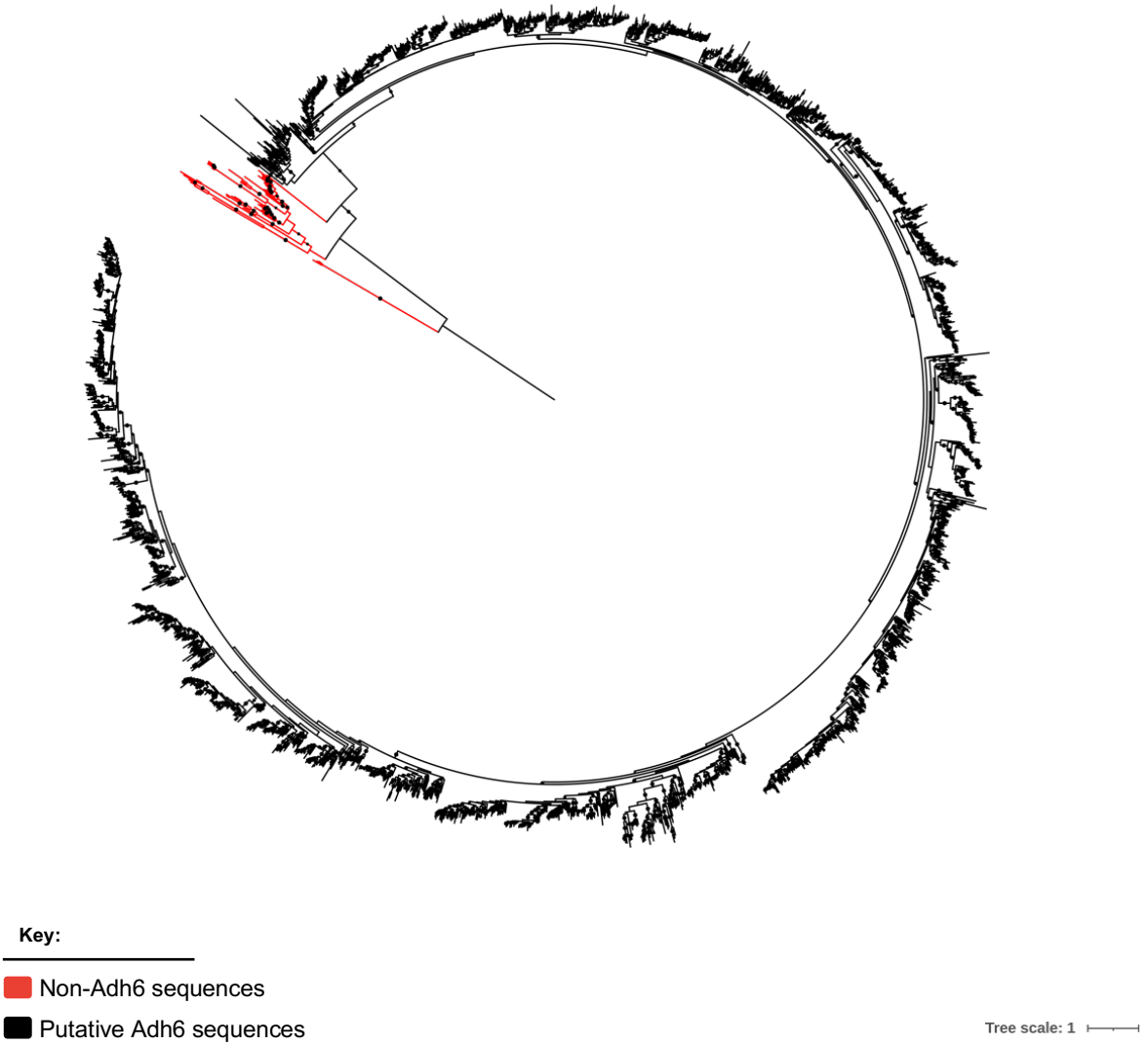


Figure A2. Preliminary Adh6 phylogeny. The preliminary tree of Adh6 was constructed with the top 5.000 BLASTp hits, using one of the *Starmerella bombicola* Adh6 sequence as query. The predicted Adh1 and Adh6 protein sequences from the W/S-clade genomes were added. Two distinct groups are observed in this phylogeny. Some protein sequences from the W/S-clade genomes, form the separate group of non-Adh6 sequences (red). This group do not cluster with the putative Adh6 sequences (black) and was removed in order to construct the final Adh6 phylogeny that is observed in Figure 3.4. Bootstrap values are indicated with black circles (>90%).