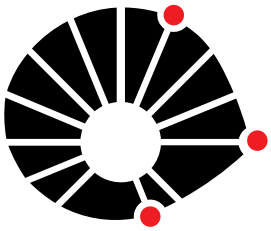


UNICAMP

UNIVERSIDADE ESTADUAL DE CAMPINAS

Instituto de Biologia

MATEUS BERNABÉ FIAMENGHI

Caracterização evolutiva e funcional de transportadores de xilose

Evolutionary and functional characterization of xylose transporters

Campinas

2023

Mateus Bernabé Fiamenghi

Evolutionary and functional characterization of xylose transporters

Caracterização evolutiva e funcional de transportadores de xilose

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

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

Supervisor: Gonçalo Amarante Guimarães Pereira

Este exemplar corresponde à versão final da Tese defendida pelo aluno Mateus Bernabé Fiamenghi e orientada pelo Prof. Dr. Gonçalo Amarante Guimarães Pereira.

Campinas

2023

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

F442e

Fiamenghi, Mateus Bernabe, 1996-
Evolutionary and functional characterization of xylose transporters / Mateus Bernabe Fiamenghi. – Campinas, SP : [s.n.], 2023.

Orientador: Gonçalo Amarante Guimarães Pereira.
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Biologia.

1. Evolução. 2. Genômica comparativa. 3. Aprendizado de máquina. 4. Açúcar - Transporte. 5. Xilose. 6. Etanol. I. Pereira, Gonçalo Amarante Guimarães, 1964-. II. Universidade Estadual de Campinas. Instituto de Biologia. III. Título.

Informações Complementares

Título em outro idioma: Caracterização evolutiva e funcional de transportadores de xilose
Palavras-chave em inglês:
Evolution
Comparative genomics
Machine learning
Sugar - Transportation
Xylose
Ethanol
Área de concentração: Genética de Microorganismos
Titulação: Doutor em Genética e Biologia Molecular
Banca examinadora:
Gonçalo Amarante Guimarães Pereira [Orientador]
Marcelo Mendes Brandão
Marcelo Brocchi
Jeferson Gross
Ricardo Cerri
Data de defesa: 05-06-2023
Programa de Pós-Graduação: Genética e Biologia Molecular

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0003-4535-8594
- Currículo Lattes do autor: http://lattes.cnpq.br/9005713927447046

Membros da Banca

Tese de Doutorado defendida por Mateus Bernabé Fiamenghi e aprovada em 05 de junho de 2023 pela banca examinadora constituída pelos doutores:

Prof. Dr. Gonçalo Amarante Guimarães Pereira - Orientador

Prof. Dr. Marcelo Mendes Brandão

Prof. Dr. Marcelo Brocchi

Prof. Dr. Jeferson Gross

Prof. Dr. Ricardo Cerri

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

Este trabalho é dedicado ao meu tio João Marcus Bernabe (Kiko)
Obrigado por todo o apoio e incentivo.
Descanse em paz.

Acknowledgements

I would like to express my deepest gratitude to my family and friends who have been a constant source of encouragement and support throughout my academic journey, especially my mom Cristina, my dad Geraldo, and my godmother Moira. Their continuous encouragement helped me to achieve this significant milestone. I am especially grateful to João Bueno, Thiago Ribas, Gustavo Bortolo for their friendship during this period, their encouragement, valuable scientific discussions and fun moments at the pub!

I would like to express my sincere thanks to my thesis advisor prof. Gonçalo, who provided guidance, counselling and provided the means for the research to be fulfilled in the lab, both experimentally and *in silico*. I would also like to thank Dr. Juliana José who provided me with invaluable guidance and support throughout my research. Her wisdom and expertise in evolution and bioinformatics has been invaluable, and without her I wouldn't have become the scientist that I am today. I am forever grateful for her help, support and friendship.

Finally, I would like to express my gratitude to the funding agencies that supported my studies, especially the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 with grant numbers 88882.329486/2019-01 and 88887.502245/2020-00) and the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), with process 2020/07918-7. Without their financial support, this research would not have been possible.

Thank you all for your support, encouragement, and guidance, and for making this achievement possible.

"Talent is a pursued interest.
In other words, anything that
you are willing to practice
you can do.
(Bob Ross)

Resumo

Fontes de energia renováveis ganharam importância para mitigar efeitos das mudanças climáticas, sem perder segurança energética. O Brasil tem tradição no bioetanol de primeira geração (1G), onde o açúcar e o melaço de cana são usados como fontes de fermentação pela levedura *Saccharomyces cerevisiae*. Embora existam várias vantagens ao uso de combustíveis fósseis, o processo 1G enfrenta críticas devido ao desmatamento e à redução da segurança alimentar. O etanol de segunda geração (2G) surgiu para resolver estas críticas e aumentar a produtividade. Produzido a partir de matérias-primas lignocelulósicas, como resíduos agrícolas, no Brasil a principal matéria-prima utilizada é o bagaço residual após a colheita da cana de açúcar. A viabilidade do etanol 2G depende da solução de desafios associados à sua produção, como a baixa eficiência da fermentação da xilose por microrganismos utilizados na indústria. A xilose, um açúcar pentose, está presente em quantidades significativas na fração hemicelulósica e não é facilmente fermentada pela levedura utilizada para a produção de etanol 1G. Portanto, é necessário modificar geneticamente estes microrganismos para utilizar xilose eficientemente. Diversas abordagens estão sendo tomadas pelos cientistas para resolver este desafio, tais como a evolução adaptativa das cepas, a engenharia genética de enzimas-chave nas vias do metabolismo da pentose e a prospecção de novos organismos sem as limitações da levedura. Um importante desafio é o transporte eficiente da xilose em levedura industrial, pois os transportadores de açúcar conhecidos têm baixa afinidade com a xilose, ou preferencialmente transportam glicose quando ambos os açúcares estão presentes. Combinamos genômica comparativa, testes de modelos evolutivos e aprendizado de máquina para estudar os transportadores de açúcar em leveduras e bactérias. A genômica comparativa se tornou uma ferramenta estabelecida para entender a história evolutiva dos organismos na busca de um fenótipo de interesse, e seu uso para entender os organismos que utilizam xilose é uma abordagem recente que demonstra promessa para a prospecção de genes de interesse. A aprendizado de máquina, por outro lado, foi utilizada com sucesso no passado para enfrentar diferentes problemas biológicos, incluindo uma ampla classificação de transportadores. Para nosso conhecimento, estas duas estratégias não foram utilizadas em conjunto para enfrentar um problema industrial e, portanto, no capítulo 1 deste trabalho visamos criar um modelo de aprendizado de máquina capaz de prever se um transportador de açúcar seria capaz de transportar xilose ou não. Este modelo foi testado em um conjunto de dados genômicos de 180 leveduras, dos quais várias famílias de transportadores de açúcar foram recuperadas e testadas contra o modelo. Quatro transportadores foram testados em leveduras e todos eles foram capazes de transportar xilose, destacando a utilidade do modelo como uma estratégia inicial de triagem para novos transportadores de xilose. O capítulo 2 apresenta um estudo evolutivo profundo do *Thermoanaerobacterium saccharolyticum*, um microrganismo extremófilo capaz de fermentar xilose e glicose simultaneamente. Muitas famílias gênicas

relacionadas ao metabolismo da xilose foram encontradas através de genômica comparativa, especialmente um transportador específico para xilose com evidência de seleção positiva perto do sítio de ligação, que pode ser explorado em estudos futuros.

Palavras-chave: evolução, genômica comparativa, aprendizado de máquina, transportador de açúcar, xilose, etanol

Abstract

Renewable energy sources have gained importance to mitigate the effects of climate change while maintaining energy security. Brazil has an established tradition in first generation (1G) bioethanol, where the sugar and molasses of sugarcane are used as sources for industrial fermentation by the budding yeast *Saccharomyces cerevisiae*. Though there are several advantages compared to fossil fuel usage, the 1G process has faced criticism due to deforestation and reduction of food security. Second generation (2G) ethanol has appeared as a complimentary process to resolve these criticisms and increase productivity. Produced from lignocellulosic feedstocks such as agricultural residues, in Brazil the main feedstock used is the residual bagasse post sugarcane harvesting. The feasibility of 2G ethanol depends on solving challenges associated with its production, such as the low efficiency of xylose fermentation by microorganisms used to produce ethanol in industry. Xylose, a pentose sugar, is present in significant quantities in the hemicellulose fraction and not easily fermented by the commonly used yeast for 1G ethanol production. Therefore, it is necessary to genetically modify these microorganisms to efficiently utilize xylose. Several approaches are being taken by scientists to resolve this challenge, such as adaptive evolution of strains, genetically engineering key enzymes on pentose metabolism pathways and prospecting novel organisms that are unincumbered by the issues of yeast. One important challenge in this field is the efficient transport of xylose into industrial yeast, as the known sugar transporters have low affinity to xylose, or preferentially carry glucose when both sugars are present. We combined comparative genomics, evolutionary model tests and machine learning powerful approaches to study sugar transporters in yeasts and bacteria. Comparative genomics has been an established tool for understanding the evolutionary history of organisms when searching for a phenotype of interest, and its use for understanding xylose-utilizing organisms and how their adaptations can be harnessed for industrial applications is a newer approach that has already shown promise for prospecting genes of interest. Machine Learning on the other hand, has been used successfully in the past to tackle different biological problems including broad transporter classification, To our knowledge, these two strategies have also not been used together to tackle an industrial problem and thus, in chapter 1 of this work we aimed to create a machine learning model capable of predicting if a sugar transporter would be capable of transporting xylose or not. This model was tested on a comparative genomics dataset of 180 yeast genomes from which several sugar transporter families were retrieved and screened against the model. Four transporters were tested in yeast and all of them were able to transport xylose, highlighting the usefulness of the model as an initial screening strategy for novel xylose transporters. Chapter 2 presents a deep evolutionary study of *Thermoanaerobacterium saccharolyticum* an extremophile microorganism capable of fermenting xylose and glucose simultaneously. Many gene families related to xylose metabolism were found through

comparative genomics, especially a specific xylose sugar transporter with evidence of positive selection near the binding site, which can be explored in future studies.

Keywords: evolution, comparative genomics, machine learning, sugar transporter, xylose, ethanol

Contents

Introduction

The continuous use of fossil fuels as an energy source has reached a worrying point, mainly concerning the consequences of greenhouse gas emissions to climate change, such as increasing global temperature and sea levels [14]. For this reason, the scientific community has searched for alternative energy sources as a means to mitigate the consequences of climate change and dependence on non-renewable resources. The study and use of renewable sources is thus essential to mitigate current and future impacts of climate change, and biofuels have gained much importance in this field [23]. Brazil's long history with biofuels can be highlighted starting in the 1970s, when a shift towards first generation (1G) biofuels such as ethanol was driven by a combination of governmental policies, economic incentives and technological advances [53]. This shift was largely due to the oil crisis of the time, when Arab countries imposed an embargo on petroleum availability, forcing the search for alternative sources of energy. This technological revolution allowed for the production of biofuels on a larger scale, and the economic incentives provided by the government encouraged the development of these alternative energy sources [5]. Cars in this period were produced to run solely on ethanol [57], but subsequent advances led to the creation of flex cars, which are able to run both on ethanol and petrol. Brazil's success in this field is internationally recognized, with many other countries looking to emulate their success in order to reduce their reliance on oil as a primary source of energy [47]. The so called 1G ethanol production in Brazil is mainly based on a process where the sugar from the sugarcane juice are used as sources for industrial fermentation by the budding yeast *Saccharomyces cerevisiae* [7, 23]. The preparation process of this feedstock is relatively straightforward, as seen in figure 1, and yeast readily utilizes the hexose sugars from this source for ethanol production by harnessing the glycolytic pathway. Interestingly, a feature of yeast that has been observed during glucose consumption is its preference for the fermentative process instead of respiration even in oxygen-rich conditions [20], in spite of the lower acquisition of ATP during fermentation. This preference for fermentation is a phenomenon still under discussion on possible reasons for its occurrence with two main ideas: the "make-accumulate-consume" (MAC) strategy, which believes that the ethanol produced from fermentation is an adaptation to hinder other organism's growth due to its toxicity, and the "rate-yield tradeoff" (RYT) strategy, which believes there is a tradeoff between fermentation and respiration due to thermodynamic constraints in ATP production or metabolic pathway constraints (due to the cost of forming intermediary products or the need to synthetize specific enzymes), as respiration can be limiting, the excess sugar can be fermented or ignored [55]. Scientists classify organisms in which the fermentation in the presence of oxygen trait is observed as Crabtree-positive [48]. Brazil's

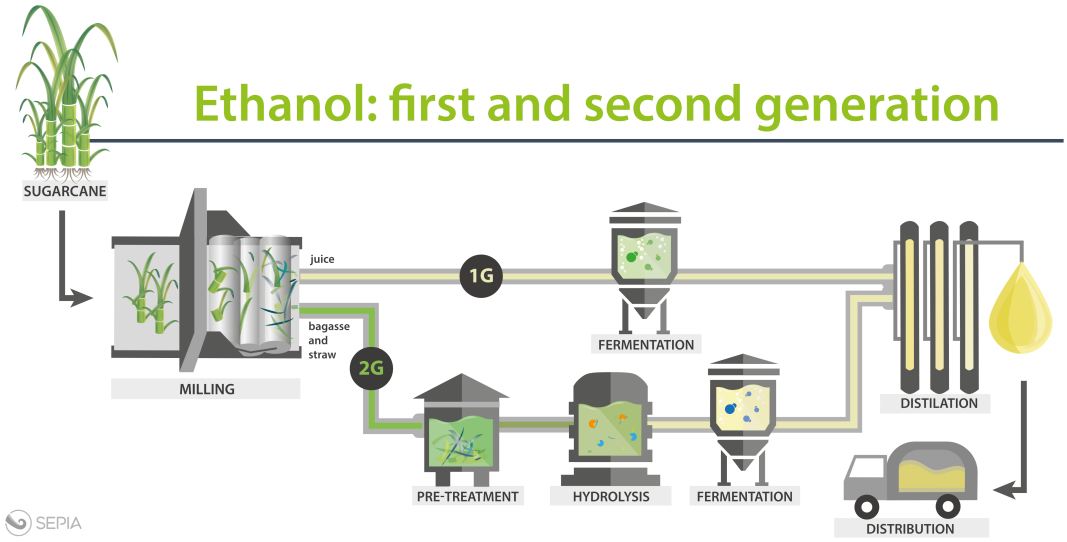


Figure 1 – First and second generation ethanol production steps. For the 1G process, after milling, the juice is sent for fermentation by microorganisms, which will use the hexose sugars present in this media and produce ethanol as a byproduct of metabolism. The 2G process, on the other hand, requires the steps of pretreatment and hydrolysis before the sugars are ready to be fermented by industrial microorganisms. Fermentation and distillation are common steps of both 1G and 2G processes, although 2G yeasts require several modifications to be able to successfully produce ethanol, as the lignocellulosic sugar composition differs from the juice sugar composition.

advantage is also credited to the manner in which the industrial process takes place. In contrast to other leaders in the ethanol field such as the USA, Brazilian process is done with an open fermentation vat, allowing contamination to occur. Over time, evolutionary pressures during fermentation allowed wild yeast strains to colonize, take over and more efficiently produce ethanol than original inoculated strains [8, 23, 46]. A few examples of these more efficient strains are CAT [65], PE [6] and BG, which have been isolated and have since been used both industrially and as chassis for laboratory improvements in research. As it is important that research replicates industrial conditions as much as possible, having these strains as a starting point for genetic manipulations is advantageous to using common laboratory strains such as CEN-PK and S288C. However, even though the 1G process brought and still brings much benefit to energy security and less emissions, some criticism has been given in the past, mainly due to deforestation for creating farmlands, burning of sugarcane before manual harvesting (a practice that has been reduced to a minimum with new legislations) and reduction of food security because of preferential planting of sugarcane instead of food crops [52].

0.1 Second generation biofuels and challenges

Second generation (2G) ethanol has appeared as a complimentary or alternative process in this field aiming to resolve these valid criticisms. Produced from lignocellulosic feedstocks such as agricultural residues, energy crops, and algae, it has gained a lot of interest due to its many benefits, such as the ability to be produced without competing with food crops for land and water resources, a production chain that can be carbon neutral or even carbon negative, the lack of need for new farmable lands, and development of novel economic chains (Wietschel et al., 2021). The lignocellulosic biomass is comprised of cellulose (30-50%), hemicellulose (25-30%) and lignin (15-20%) [23, 12], as shown in 2. In Brazil, the main feedstock used for 2G ethanol is the residual bagasse that is left on fields post sugarcane harvesting, however, novel approaches have been gaining traction, such as the use of energy crops, for instance Energy Cane [23, 29] and Agave [60], in infertile lands such as the northeastern Sertão, to better utilize these impoverished regions with the benefit of economically developing the resident communities [1, 23]. The 2G process is comprised of the following steps, graphically shown in figure 1: 1. Pretreatment, during which the sugars within the lignocellulosic biomass are released into the media, through chemical (such as acid or alkaline pretreatment, or oxidative delignification), physical (milling, pyrolysis, or microwaving), physicochemical (such as steam ammonia fiber or CO2 explosion) or biological (fungal, bacterial, and enzymatic biopulping) methodologies. This is a necessary step to break down the complex structure of the biomass into accessible components for the enzymes in the next stage [37]. 2. Enzymatic Hydrolysis, where enzymatic cocktails are used on the pretreated biomass to break down the cellulose and hemicellulose portions into simpler sugars such as glucose and xylose. 3. Fermentation, the resulting sugar solution is then fermented by microorganisms, predominantly the yeast *S. cerevisiae*, that convert the sugars into ethanol. 4. Distillation, the fermented solution is then distilled to separate the ethanol from the remaining water and other impurities. However, the feasibility of 2G ethanol depends on resolving some challenges associated with its production, such as the formation furfural HMF, acetic acid and phenolic compounds, which are fermentation inhibitors [56]. Another important hurdle is optimizing the enzymatic hydrolysis step with cheaper and more efficient enzymes, capable of activity at lower temperatures, in a way to better couple this step with fermentation. Regarding fermentation one important challenge to overcome is the low efficiency of xylose fermentation by microorganisms used to produce ethanol in industry. Xylose, a pentose sugar, is present in significant quantities in the hemicellulose fraction and is not easily fermented by the commonly used yeast for 1G ethanol production, which is unable to natively utilize xylose as a sole carbon source [67]. Therefore, it is necessary to genetically modify these microorganisms to efficiently utilize xylose.

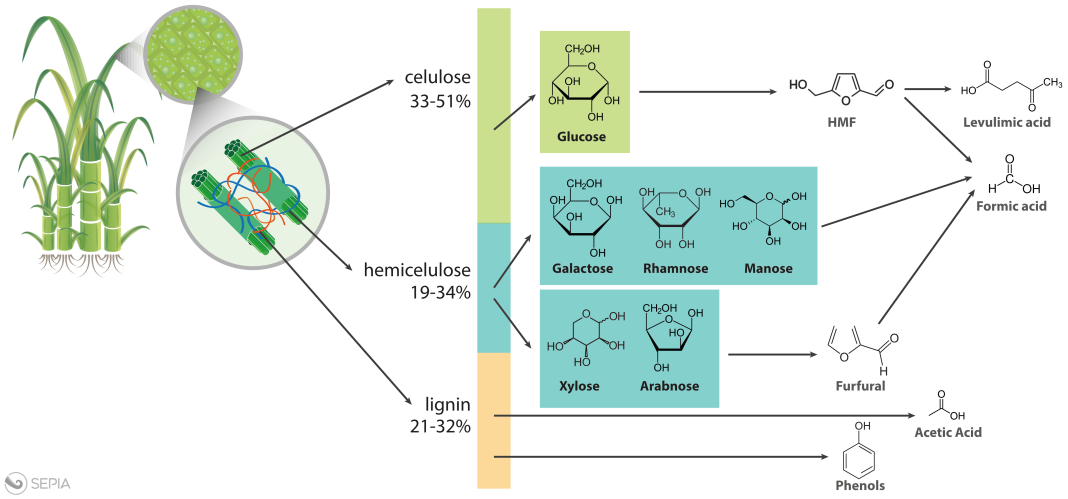


Figure 2 – Composition of biomass used for 2G ethanol, with sugars obtained from each fraction and fermentation inhibitors produced during pre-treatment of this biomass.

0.2 Xylose metabolism and challenges

There are a few pathways related to the consumption of xylose, but two are the most commonly studied. The oxidoreductive pathway follows the conversion of xylose into xylitol through the enzyme xylose reductase (XR), followed by a conversion of xylitol into xylulose by xylitol dehydrogenase (XDH), while the isomerase pathway has the immediate conversion of xylose into xylulose through the enzyme xylose isomerase (XI). Both pathways have in common the subsequent conversion of xylulose into xylulose-5-phosphate by xylulokinase (XKS) which follows to the pentose phosphate pathway, and have inherent inefficiencies associated to them [75, 10]. The oxidoreductive pathway suffers from a redox imbalance because XR preferentially uses NADPH over NADH, while XDH uses NAD^+, causing NADH to accumulate. As NADH cannot be fully oxidized through the respiration process, this leads to the accumulation of xylitol and reduces ethanol yield [76]. Even though $NADP^+$ may be slowly reduced to NADPH on the Pentose Phosphate Pathway, NADH is usually oxidized on the mitochondrial respiratory chain, which poses an industrial problem with fermentations being done on almost anaerobic conditions due to the large volume in the bioreactors [23].

The isomerase pathway on the other hand, although being much simpler and not being constrained by redox unbalances, has low throughput and is not native to *S. cerevisiae*, with the additional problem of many heterologous XI that have been prospected are not functional in this yeast [27, 38, 69]. Both pathways are shown in a simplified manner in figure 3.

Another important challenge to be overcome, and the one that was the focus of this work, is the internalization of xylose itself, which is done through sugar transporters.

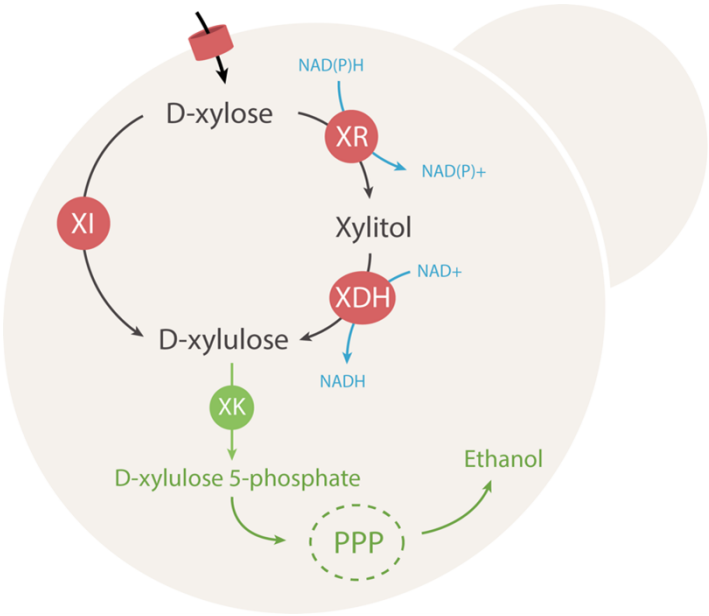


Figure 3 – Simplified main pathways of xylose metabolism to ethanol in yeast. The glycolytic pathway following the Pentose Phosphate Pathway (PPP) is omitted for brevity.

Yeast presents a preferential consumption of glucose and only switches its metabolism to other available carbon sources when glucose is depleted, a preference that is justified by the lower metabolic costs associated with a direct entrance into the glycolytic pathway [25] and higher affinity for glucose of the hexose transporters [34, 24]. Xylose uptake in the presence of glucose can be improved by the overexpression of pentose transporters [66], however robust xylose transporters are rare, and the already described proteins are strongly inhibited by glucose [40], making direct evolution steps necessary to improve consumption when using these transporters [43, 63, 71]. Prospecting and screening novel transporter candidates that may be able to carry xylose or other pentoses in industrial organisms is also extremely relevant and one of the main objectives of this work.

0.3 Sugar transporters and challenges

Regarding more deeply sugar transporters, they represent a wide group of proteins branched in many families and superfamilies, for instance the Major Facilitator Superfamily (MFS) [59] and the ATP-Binding Cassette Superfamily (ABC) [26] each having different transport mechanisms and evolutionary origins. Figure 4 demonstrates the overall structure of both transporter classes and their mechanism of action.

MFS transporters are highly conserved, usually having 12 transmembrane loops and are believed to have diverged from an initial gene duplication prior to their separation in subfamilies. MFS transporters can be of three main types: uniporters, symporters, and antiporters. These three types carry their solutes following the diffusion gradient,

however uniporters carry only the solute across the membrane, while symporters also carry a second molecule simultaneously on the same direction and antiporters require that another molecule be transported on the opposite direction [74].

When analyzing these groups through an evolutionary scope, substrate carried and the phylogenetic history of a transporter family have a certain degree of correlation [54, 28], but due to the high degree of promiscuity of sugar transporters in their sugar transport capacity, it is difficult to predict specificity and efficiency through phylogenetic history alone [39]. MFS sugar transporters are usually under purifying selection, meaning that changes in amino acids are usually detrimental to the organism and removed by natural selection. Considering all these points, the evolution of sugar transporters must also be considered under broader yeast adaptations into their habitats and events such as the Whole Genome Duplication (WGD), an event which occurred approximately 100-150 million years ago which likely involved mating between two different ancestral yeast species followed by a doubling of the genome [49, 73].

Gain and losses of transporters, as well as specialization might be indicatives of niche adaptation, as gene expansion and duplication of transporter families has been reported as a means of increasing efficiency. Examples of these occurrences are the Hxt transporters in crabtree-positive yeasts [45, 20], which evolved in a dose-dependent manner (having more copies of a transporter directly correlates with the amount of sugar that can be internalized) after WGD, the KHT and HGT families in *Kluyveromyces marxianus* [22], and the Snf3 or Rgt2 which lost the ability to uptake sugars and instead act as sugar sensors that influence gene expression [21]. Though maintaining more copies of a transporter in the genome may be energetically hindering [70, 2], gene duplication can facilitate sugar uptake in high sugar environments and provide an advantage in low sugar conditions [11, 35]. Gene duplication can also provide the opportunity for positive selection of novel, beneficial functions, allowing one duplicated copy to potentially gain affinity for an alternative sugar.

Other factors also need to be considered when attempting to predict the substrate selectivity of a transporter. In particular, the amino acid sequence of the transporters provide further insight into the physiology and structure of the resulting proteins (Young et al., 2014). As highlighted in the literature, conserved amino acid motifs may be key in conferring preference for certain sugar sources [24]. ABC transporters on the other hand carry their solutes against the concentration gradient, needing energy expenditure to transport their substrate. These transporters can function as importers or exporters and are ubiquitous to most groups, however, importers are not found in most eukaryotic organisms, and are predominant in bacteria, suggesting a loss of function during evolutionary history [15]. ABC transporters are highly conserved even between eukaryotes and prokaryotes in some of their domains: the two nucleotide binding domains (NBDs),

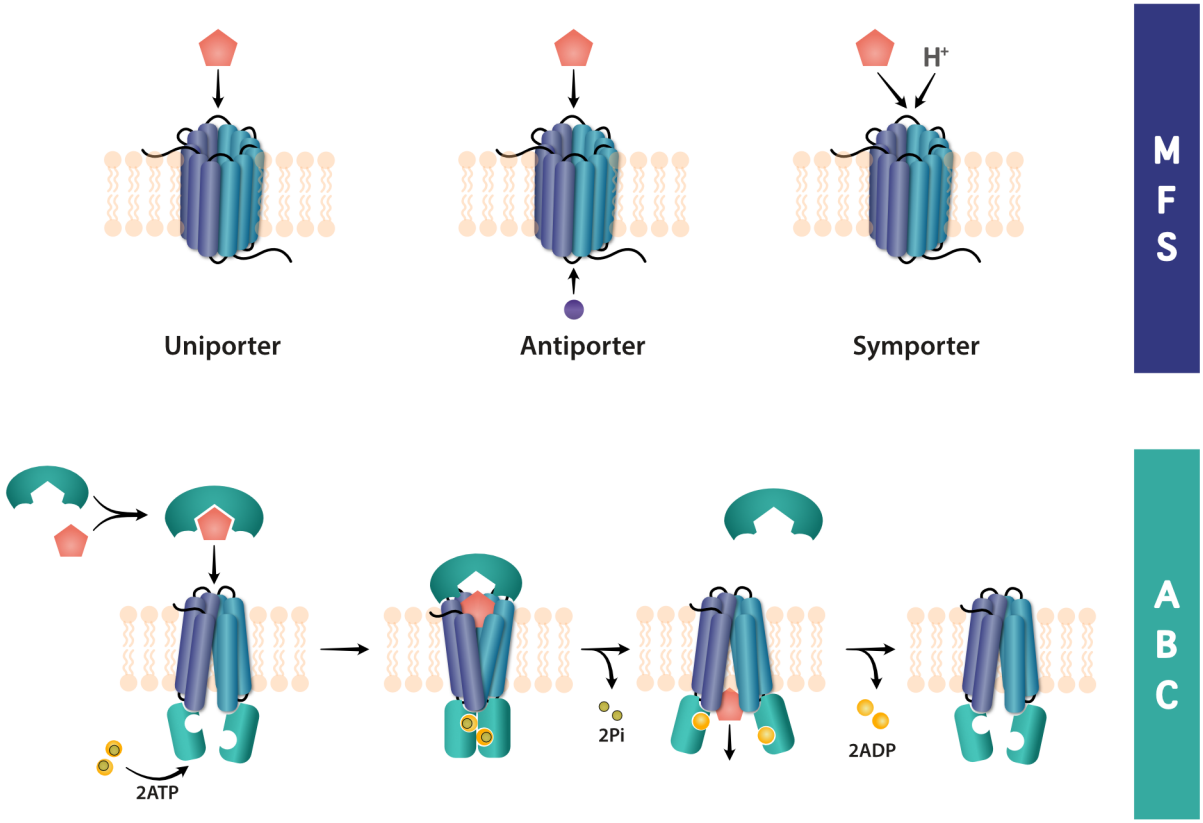


Figure 4 – Schematic representation of MFS transporters and ABC importers. MFS transporters have 12 transmembrane loops (shown in purple and green) with three transport mechanisms: uniport, where the sugar molecule (shown in red) is transported following the concentration gradient; antiport, where the sugar is transported while another molecule (purple circle) is transported on the opposite direction; symport, where the sugar is transported together with an ion. ABC importers have 6-10 transmembrane loops (shown in purple and green), an internal ATP binding domain (in green, below the transmembrane region) and a sugar binding domain (in green, above the transmembrane region), the sugar binding domain is specific to its sugar and after binding brings the solute close to the transmembrane domain. The transmembrane domains open after ATP binding and hydrolysis in the ATP binding domain, allowing the sugar to be internalized.

responsible for binding ATP and making the mechanism work and the two transmembrane domains (TMDs), that open and close in response to the binding of ATP [15]. ABC importers, contrary to what is seen for MFS transporters, are specific to their substrate carried, and usually organized in bacterial operons, with an additional extracellular domain besides the TMDs and NBDs, which is responsible for binding the sugar molecule and bringing it to the other domains for import [17]. A few examples of these importer systems are the MsmEFGK, responsible for transporting raffinose and stachyose in *Streptococcus mutans*, and xylFGH, responsible for xylose transport, in *Escherichia coli*.

0.4 Finding solutions in yeast

Comparative genomics is a way to analyze and study different organisms based on their genomic similarities and differences, focusing or not in a phenotype of interest. By comparing organisms with some traits against other organisms of the same group without said traits, it is possible to find evolutionary footprints that may have contributed to these differences [16]. This approach allows researchers to study evolutionary relationships between organisms and to gain insights into the molecular mechanisms that underlie biological diversity, shedding light on the evolution of life on Earth, as by comparing the genomes of different organisms, it is possible to trace the evolutionary relationships between species and identify the genetic changes that have contributed to the development of new traits and adaptations. This information can help us to better understand the origins of biodiversity and the mechanisms that drive evolution, as well as provide clues to the genetic basis of complex traits and diseases. In addition to these phylogenetic methods, machine learning has also gained traction to predict function or classify proteins in biological datasets. Machine learning is a term used to mean the process of using algorithms and statistical models to find patterns and insights from data sets. These algorithms are usually divided into two main classes: supervised learning and unsupervised learning algorithms [4]. Unsupervised learning deals with unlabeled data, where the training dataset contains only input features without any corresponding output labels. The goal of unsupervised learning is to discover inherent patterns, structures, or relationships in the data without prior knowledge of the outcomes. Supervised learning on the other hand is a type of machine learning where the algorithm is trained on labeled data. In this approach, the training dataset consists of input data (features) and corresponding desired output labels. The goal of supervised learning is to build a model based on the features that can predict the labels for new, unseen data. This is also the methodology implemented in this work.

Work has been done with machine learning on predicting if an RNA molecule is a coding RNA or lncRNA [13], prediction of deleterious genetic variants [58] and even for comparative genomics [31] or metabolic system evolution [36]. For transporter proteins specifically, studies have been published to classify sequences into transporters and non-transporters [3] or to find the correct superfamily for a specific transporter [44, 30], predict substrate specificity [50] or understand and predict structural conformations [51]. However, an attempt to predict xylose transport capacity had not yet been reported, although it could be an interesting strategy for initial screening of xylose transporter candidates.

As mentioned before, the ability to predict the transported sugar through phylogenetic history alone for MFS transporters is extremely challenging, nevertheless, evolutionary studies based on comparative genomics can tell us a lot about a gene's capabilities and analyzing the phylogenetic history of species with a phenotype of interest for industrial applications is a newer strategy that has been gaining traction on the

scientific community [72, 62, 9]. Thus, one of the main objectives of this work was to detect selective clues in xylose fermenting or consuming species of the Saccharomycotina clade that may have been beneficial for this phenotype, focusing on finding novel xylose transporter candidates that may be used for genetic engineering in industrial yeast strains. Coupled with this evolutionary methodology, a supervised machine learning algorithm was created attempting to find patterns in known xylose-transporting sugar transporters and predicting from the evolutionary dataset which transporter candidates would be more suited for wet-lab validation when compared to one of the best xylose transporters known in literature, GXF1 [41]. This double approach has not been done before and with continued use of the machine learning model, it will be possible to more easily predict the capability of a transporter protein to carry xylose, as well as understand the molecular and evolutionary mechanisms that decide if a transporter is able to transport xylose or not. This work can be found on chapter 1.

0.5 Finding solutions in bacteria

One other approach that has been gaining traction is the search for alternative organisms to substitute industrial yeast as microbial cell factories in the 2G industry. Although this may be possible in the future, there is still much work to be done to equal the advances done in yeast. Nevertheless, understanding these organisms is important to find adaptations and candidate genes that can help us to better understand the different strategies and underlying mechanisms of xylose consumption, and explore new candidate genes that may also be used in current yeast cell factories. One organism that should be highlighted for industrial applications is *Thermoanaerobacterium saccharolyticum* (*T. sac*), an extremophile microorganism from the firmicutes group, found originally in hot springs and geysers [42]. This organism can naturally consume xylose, cellobiose and many other sugars of interest, grow in insoluble hemicellulose [33, 18], is easily genetically manipulated [32] and has been engineered to ferment xylose to ethanol at high yield [64]. More interestingly is its ability to co-ferment xylose and glucose, an extremely rare phenotype of high industrial interest [64]. Additionally, microarray data has been published for *T. sac* while fermenting xylose and some other interesting conditions [19]. However, evolutionary studies of this organism have not been done.

As mentioned before, bacteria harbor ABC importers as well as the common MFS transporters. Due to the necessary energy expenditure for assimilation of the solutes when using an ABC system, we hypothesized that these transporters would show evidence of positive selection in this portion, due to the great specificity seen to the substrate carried, and due to the nature of transport against the concentration gradient. Although untested, we speculated that these transporters would be expressed after much of the carbon source had already been depleted as a competitive advantage against other organisms in the

environment and more specific xylose transporters would have been positively selected. On chapter 2 a deep comparative genomics study of *T. sac* is presented, comparing against other firmicutes, and searching for genes that may be related to xylose metabolism, especially on sugar transporters. Among many discussions regarding selection pressures and gene duplications in many genes that may be related to xylose metabolism, positive selection was found on an extracellular binding protein, similar to xylF of *E. coli*, close to the sugar binding site, highlighting that this *T. sac* operon may be interesting for industrial applications and should be further explored.

1 Machine learning and comparative genomics approaches for the discovery of xylose transporters in yeast

1.1 Abstract

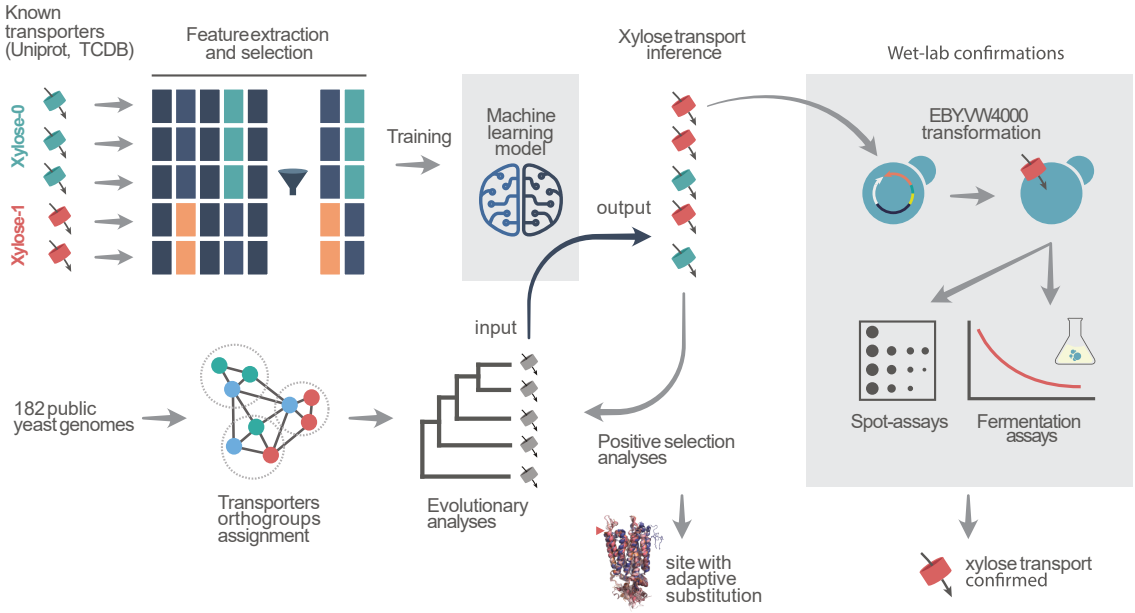
Background: The need to mitigate and substitute the use of fossil fuels as the main energy matrix has led to the study and development of biofuels as an alternative. Second-generation (2G) ethanol arises as one biofuel with great potential, due to not only maintaining food security, but also as a product from economically interesting crops such as energy-cane. One of the main challenges of 2G ethanol is the inefficient uptake of pentose sugars by industrial yeast *Saccharomyces cerevisiae*, the main organism used for ethanol production. Understanding the main drivers for xylose assimilation and identify novel and efficient transporters is a key step to make the 2G process economically viable.

Results: By implementing a strategy of searching for present motifs that may be responsible for xylose transport and past adaptations of sugar transporters in xylose fermenting species, we obtained a classifying model which was successfully used to select four diferent candidate transporters for evaluation in the *S. cerevisiae* hxt-null strain, EBY. VW4000, harbouring the xylose consumption pathway. Yeast cells expressing the transporters SpX, SpH and SpG showed a superior uptake performance in xylose compared to traditional literature control Gxf1.

Conclusions: Modelling xylose transport with the small data available for yeast and bacteria proved a challenge that was overcome through different statistical strategies. Through this strategy, we present four novel xylose transporters which expands the repertoire of candidates targeting yeast genetic engineering for industrial fermentation. The repeated use of the model for characterizing new transporters will be useful both into finding the best candidates for industrial utilization and to increase the model's predictive capabilities.

Keywords: Xylose, Xylose transporter, Machine learning, Feature selection, Pentose metabolism, Industrial biotechnology

1.2 Graphical Abstract



1.3 Background

For the last few decades, the scientific community has expended efforts to find cleaner energy alternatives to the fossil-based matrix as a means to mitigate the consequences of climate change from the use of said fossil fuels. One strategy is the use of biofuels produced from lignocellulosic biomass, also called second-generation (2G) biofuels. This strategy is desirable, as lignocellulose is found in the cell walls of all plants and allows many different matrices to be used industrially [1,2,3,4].

Plant cell walls comprise mainly cellulose (30–50%), hemicellulose (25–3%) and lignin (15–20%) [5]. The 2G process involves breaking down these main saccharide fractions into their monomers, predominantly glucose and xylose, that can then be metabolized by microorganisms into different bioproducts, and in the context of biofuels, bioethanol [6,7,8]. 2G biofuels appear as a promising driver on energy security due to it not competing directly with the food industry and not needing more plantations to achieve energy security [8].

Xylose consumption follows two pathways: the first, deemed the oxidoreductive pathway, comprises a conversion of xylose into xylitol through the enzyme xylose reductase, followed by a conversion of xylitol into xylulose by xylitol dehydrogenase. The second, called the isomerase pathway, comprises a one-step conversion of xylose into xylulose by xylose isomerase. Both pathways then have a conversion of xylulose into xylulose-5-phosphate by xylulokinase which then follows the pentose phosphate pathway [9].

The main organism used industrially for these biotechnological applications is

the yeast Saccharomyces cerevisiae due to its resistance to inhibitors, high product yield, and ease of manipulation [9, 10]. However, its utilization of xylose naturally is lacking, requiring genetic engineering steps to insert one of the two xylose metabolism pathways to ethanol. Although yeast strains with these pathways are already used extensively, the challenge of xylose consumption remains, related to the cofactor imbalance on the oxidoreductive pathway, the need to further engineer or evolve exogenous xylose isomerases on the isomerase pathway, or inhibition of the xylose pathway by glucose due to sugar phosphorylation mechanisms [9, 11,12,13,14].

Regarding xylose transport, *S. cerevisiae* has many hexose transporters that are also capable of transporting xylose, such as the Hxt family of transporters and Gal2 [15,16,17]. Many xylose transporters have been found in other yeast species and considered candidates for industrial use by engineering S. cerevisiae, such as Sut1-3 and Xut1 from *Scheffersomyces stipitis* [18], Gxs1 and Gxf1 from *Candida intermedia* [19] and XylHP from *Debaryomyces hansenii* [20]. Details of their kinetic properties in xylose and glucose have also been described [15, 21,22,23,24]. Besides yeasts, one of the most studied and known xylose transporters is xylE from *Escherichia coli* [25,26,27].

Xylose consumption rates decrease when coupled with glucose due to competition of these sugars by endogenous transporters for access to the transport system, where first the organism depletes all the hexoses in its media, and only then slowly metabolizes xylose [15, 28, 29]. Even though many xylose-transporting proteins have been described in literature, this inefficient consumption pattern remains and what defines the ability to transport xylose is not entirely understood. Also, sequence, evolutionary or chemical interaction characteristics (hereafter discussed as features) behind transport capacity are still not completely understood [27], as much variability on transport capacity, velocity and affinity is seen, one example is the sugar transporter Gxf1, which shows an efficiency shift at certain sugar concentrations [23].

Many studies have been done to describe new transporters from new species [19, 20, 30,31,32,33], engineer hexose or pentose transporters for better efficiency through genetic engineering and directed evolution [28, 34,35,36,37,38,39,40], develop transporter testing yeast strains [16, 30, 41], resolving crystallographic structures coupled with xylose [25], but the main genomic drivers for xylose affinity, such as adaptive evolutionary signals (e.g. positive selection and convergent evolution), structural affinity and relations between the key residues already described as important for transport have not been found. This is in part due to xylose transport not having a single structural motif indicating its trait and no known specific transporters, even though many amino acid sites for different transporters have been described to be key for xylose affinity [15, 40, 42]. Also, as transporters have evolved in a multi-genic strategy (gene duplication, resulting in multiple sequences coding for the same protein) as an evolutionary solution to increase throughput and adaptation

[43], this makes it harder to choose, test and find the best candidates for industrial purposes. Understanding these kinetic dynamics is also desirable for better rational engineering of yeasts. One novel promising approach has been to understand the evolutionary history of xylose consuming yeasts compared to non-consumers, finding genomic adaptations that may have arisen in response to the need of using xylose [44, 45]. A similar genome-wide comparative genomics study searching for adaptations in key xylose utilization pathway was previously described [46]. A similar approach using comparative genomics focused on the phylogenetic structure was used to prospect and choose novel xylose transporter candidates from *Candida sojae* [47].

The use of machine learning models to classify and predict has been previously applied to transporters as a means to separate and differentiate functional classes and families [48, 49], however due in part to the many classes in which transporters fall into, the models often lack accuracy and precision. Simpler methods, such as sequence homology, topological comparison or sequence profiling have been used before to describe different proteins, including sugar transporters [48, 50,51,52,53,54], but a unified process that weights each methods' importance has not been described. The goal of this study was to cross sequence pattern information by extracting different features from known annotated xylose transporters in yeast or bacteria with past evolutionary adaptations via comparative genomics of 182 yeast genomes as an attempt to describe what genomic elements define if a sugar transporter is capable to transport xylose or not. A classification model was created and successfully used against sugar transporter families from the dataset to find potential xylose transporters. These candidates were then characterized by growing yeast expressing these candidate genes on a set of different sugars. Finally, the structure of each of these four transporters was modelled and their docking pose coupled with glucose and xylose was compared against the crystallographic structure of the known symporter from *E. coli* xylE.

In this work, we believe another step was given on facilitating the search for xylose transporters and understanding what the main drivers for xylose affinity are, while presenting a model that can already help to choose the most likely xylose-transporting candidates to take on for wet-lab work, and that with further use will become even more reliable.

1.4 Results

1.4.1 Sugar transporters from 182 yeasts cluster in 4 families

We selected 182 genomes (Additional file 1: Table S1) from the Saccharomycotina clade available for download in NCBI to try and understand the evolutionary history and adaptations of different yeasts that conferred an ability or not to ferment

or consume xylose (manuscript in preparation). Orthofinder [55] analysis followed by recovery of families of interest through BLAST with known xylose transporters as baits revealed that sugar transporters grouped into 4 orthogroups: families 9, 10, 1180 and 7608, containing 1298, 1293, 204 and 8 genes, respectively. Full protein sequences for each family are available as Additional file 4: File S1.

1.4.2 Training and testing dataset

The dataset for model selection comprised sugar transporters for fungi and bacteria as annotated and registered in Uniprot [56] and on TCDB [57]. Xylose transporters were carefully screened from these data, and due to insufficient proteins with this function, a literature search was done to increase their number. In total, 396 proteins, from which 25 were able to transport xylose, had their amino acid sequence retrieved and were used for machine learning (Additional file 2: Table S2 is given with the gene name, Uniprot ID and publication describing xylose transport). The data were split into training and test sets using scikit-learn's [58] train_test_split.

From the more than 30,000 features extracted for the sequences, 13 were defined by the model as most important, from which 2 were impactful for a xylose transport capacity classification (Xylose-1), and the other 11 for an inability to transport xylose (Xylose-0) (Fig. 1a). Most of these features are derived from profile-based descriptors, these include the Position Scoring Matrices Features (PSSM), which indicate patterns of different sequences and scores each amino acid according to its position on the sequence, and are useful for predicting function of sites or classifying residues [59], the two features that drive the prediction to xylose-1 and the custom Hidden Markov Model (HMM), which similarly to PSSM calculates and scores sequence position, sites and patterns given other known or similar sequences, that was extracted from the non-cytoplasmic domains of the sequences, the latter also being the feature with the most impact. Other important features are related to relative mutability (DAYM780201) [60], residue volume and its consequence for the final protein conformation (BIGC670101) [61], modelling possible ligand–target interaction (scl5.2lag.5) [62] and adding more information, such as hydrophobicity in relation to near residues, to the amino acid composition (Pc1.c) [63].

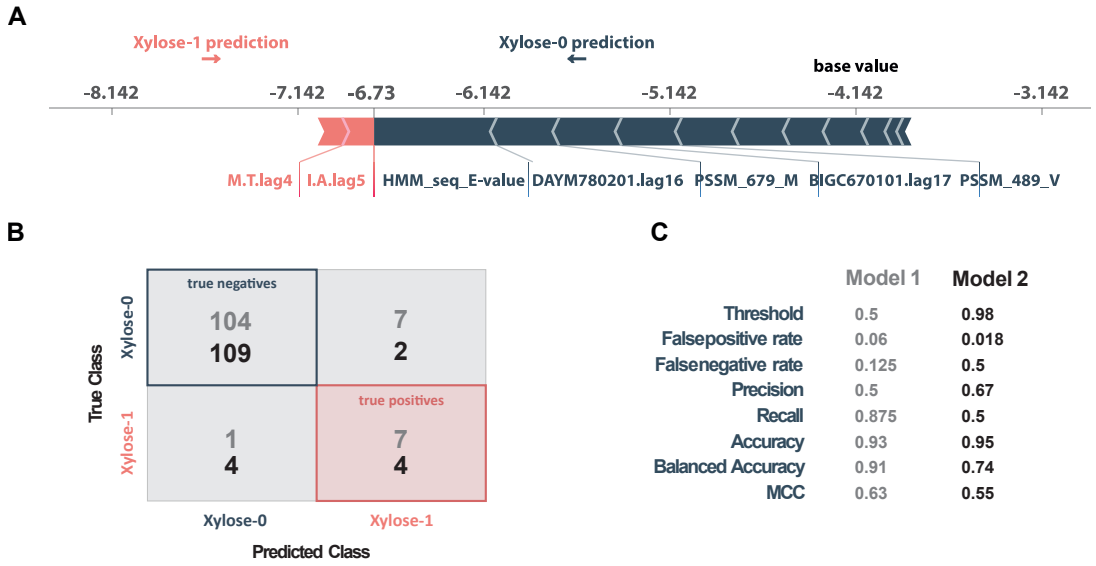


Fig 1. Graphical representations of machine learning model against the dataset. **a** Force-plot of most important features as calculated by Recursive Feature Elimination by Cross-Validation with XGBoost. Features highlighted in red are responsible for driving the final prediction of a sample into the positive category (A probable xylose transporter) while features in blue drive the prediction into the negative category (A non-xylose transporter). The base value represents the average prediction for the samples, while the size of the feature represents its impact (higher or lower importance). **b** Common metrics used to evaluate a model, the grey values correspond to the base threshold model and blue to the altered threshold. **c** Confusion matrix showing the results of predictions against the test data

Due to the dataset imbalance (only 25 out of 396 were xylose-transporting proteins), statistical oversampling techniques were implemented to reduce this imbalance. Standard classification metrics were done to evaluate the model, such as receiver operating characteristic (ROC) and precision-recall graphs, in addition to the confusion matrix which allow visualization of the absolute number of samples in each class correctly or incorrectly predicted by the model. The ROC curve showed higher increments of true positive rate than of the false positive rate, which means that the model efficiently classified the positive samples from the testing dataset (AUC=0.95 for both classes) without losing much precision. Similarly, the precision-recall analysis showed an average precision of 0.73. However, due to the initial imbalance against the positive class, and to reduce overfitting issues arising from oversampling, we sought to remove this bias by analysing the data more attentively, modifying how results were judged and giving more weight to precision. As the default model (Model 1) classification threshold is 0.5 to assign a sample to each class (0–0.49 as negative; 0.5–1 as positive), we manually edited (Model 2) so that only samples with prediction probabilities of 0.98 or higher were classified as xylose transporters. At the cost of classification power for true xylose transporters (lower recall), we were able to

almost nullify false positives for this class and thus increase precision and decrease the false positive rate (Fig. 1b).

1.4.3 Choosing transporter candidates from the comparative genomics dataset

Four transporter families were returned during our phylogenomics analysis by searching the 182 yeasts dataset families against know sugar transporters with xylose-transporting capacity (XUT1, GXF1, GXS1, HXT7, Xylhp, XUT3, xylE, Cs3894, Cs4130) through BLAST and the MFS HMM from PFAM database [64]. All sequences from these families underwent feature extraction as done for the training and testing dataset and were tested against the model with altered baseline threshold. 25 sequences were predicted as xylose transporters, from which four sequences were chosen to be tested experimentally: Spaxylofer2423 (SpX), Spagorwiae6242 (SpG), Spahagerda5424 (SpH) and Suglignoha2156 (SuL), from *Spathaspora xylofermentans*, *Spathaspora gorwiae*, *Spathaspora hagerdaliae* and *Sugiyamaella lignohabitans* species, respectively. These sequences were chosen as three of them are from the known xylose fermenting *Spathaspora* genus, and the *Sugiyamaella lignohabitans* species, which is also known to consume xylose while not being part of the fermenter's clade [46]. SuL was identified as the HXT5/HTX6 hexose transporter from *Sugiyamaella lignohabitans* [65], SpX, SpH and SpG as HXT2 from Spathaspora sp. [66] or HXT5 from *Candida subhashii* [67], through BLAST search.

All 25 sequences were from fam10. Interestingly, known xylose transporters such as Gxf1 (Caninterme1096), Cs4130 (Cansojae5099) and Cs3894 (Cansojae522) from *Candida intermedia* and *Candida sojae*, respectively, were also part of fam10, which highlights the potential of homologs in other yeast species that are seldom explored. Caution was taken when the model had not displayed these known xylose transporters in its output, however, on closer inspection of the prediction probabilities, this happened due to the increased restriction on the classification (a 0.95–0.96 threshold would have included them).

Additionally, five of these 25 sequences were found to have positive selection evidence on one codon, at protein alignment site 856, which by Interpro analysis and posteriorly by 3D modelling, was observed to be positioned on the extracellular non-cytoplasmic domain of the first helix. This result indicates that these proteins had amino acid substitutions potentially functioning as adaptations related to the xylose fermenting phenotype through their recent evolution and thus emphasizes the importance of these sequences on xylose metabolism. Due to the size of fam10 and the heterogeneity of transporters contained in it, the alignment on this site was mostly indels for most species, however, many interesting patterns were found at the positively selected codon. Firstly, being part of the non-cytoplasmic region, this site was contemplated on the HMM feature,

which was also the most impactful for the machine learning model. Secondly, as can be seen in Fig. 2, the four previously chosen candidates for experimental validation have the same amino acid (valine) at the positively selected site. While this might be expected for the three *Spathaspora* candidates and explained by it probably being an adaptation inherited from their common ancestor during speciation, the Sugiyamaella transporter also contains valine at this site while also having diverged from the clade containing *Spathaspora* much earlier during these yeasts' evolutionary history. This pattern might indicate convergent evolution at this site.

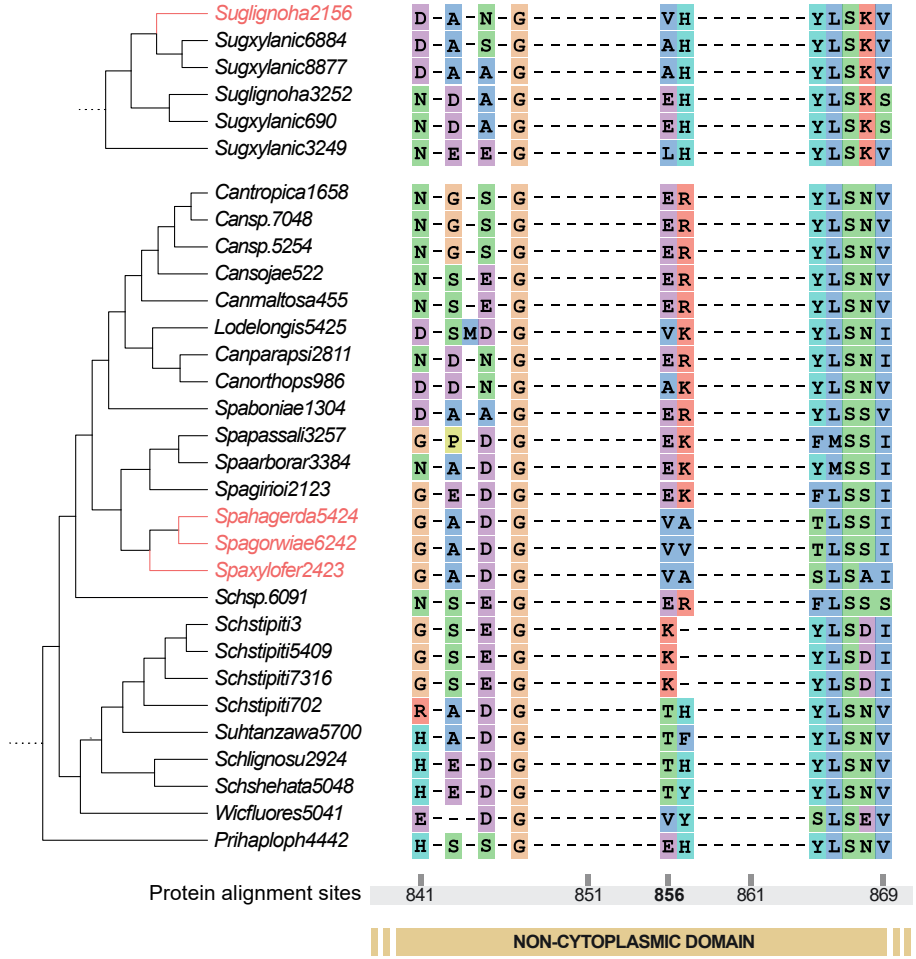


Fig 2. Snippet of fam10 phylogeny transformed into a cladogram for visualization purposes, coupled with the alignment around the site found under positive selection by MEME. In red are the transporters chosen for further characterization. Bootstraps are not shown as all of them on these clades were over 80

1.4.4 Evaluation of chosen transporters in different sugars

The substrate uptake capacity from these four sugar transporters was evaluated in the strain EBY_Xyl1, a modified yeast strain derived from EBY.VW4000 [16] lacking most of its hexose sugar transporters, rendering it unable to grow on most sugars except maltose, and engineered with the xylose oxidoreductive pathway genes. The four

transporters were codon-optimized for expression in S. cerevisiae (Additional file 5: File S2), assembled with the promoter and terminator sequences of the TDH1 gene from the glycolytic pathway and cloned into the multi-copy vector pRS426. The xylose-facilitator GXF1 from C. intermedia was cloned in the same manner as the four candidates and used as a positive control for xylose transport, since this transporter is one of the best heterologous xylose transporters described in literature [23, 68].

We analysed the substrate range of EBY_Xyl1 mutants carrying the specified sugar transporters using six different sugars on solid culture medium—2 % maltose (control), mannose, fructose, glucose, and galactose. The transformants were grown for 24 h to the exponential phase on Maltose and spotted in tenfold serial dilutions onto solid culture medium. All transporters, except for SuL, were able to confer growth of EBY_Xyl1 on all sugars, indicating a substrate promiscuity commonly seen for sugar transporters. However, SuL was especially surprising as growth in fructose, mannose and glucose was almost non-existent (Fig. 3a).

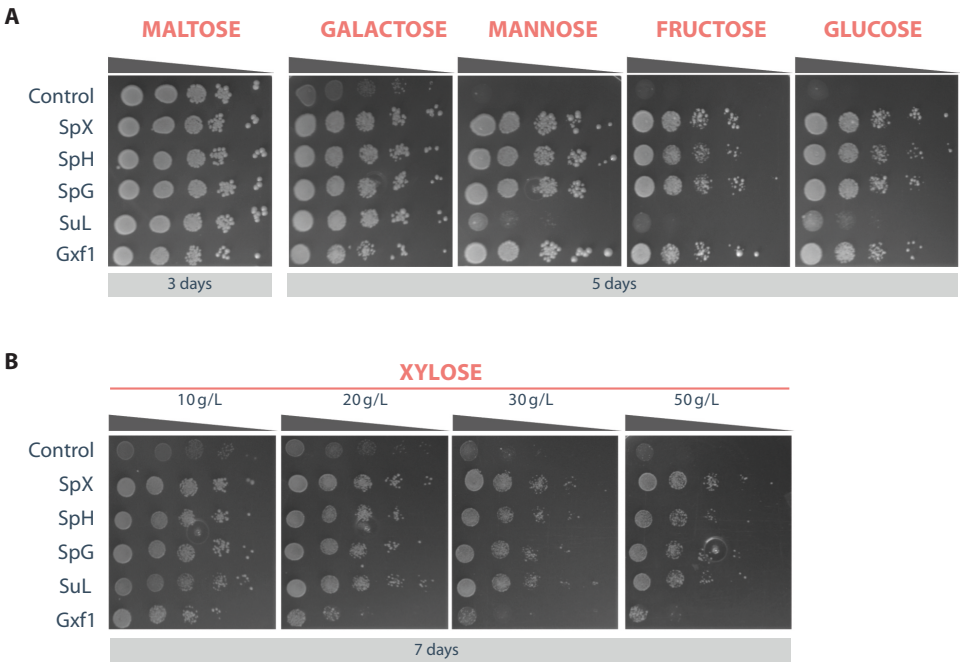


Fig 3. Spot-assay of EBY_Xyl1 carrying each of the indicated transporters and growing in **a** different sugars and **b** different concentrations of xylose. Initial OD600 was settled at 1 before the tenfold serial dilution. Plates were incubated in 30 °C. All experiments were performed in triplicate

The growth of EBY_Xyl1 mutants carrying transporter genes and Gxf1 as positive control were also compared in solid medium with 1, 2, 3 and 5% of xylose as the sole carbon source. All transporters were able to confer growth in all sugar concentrations, and the four transporters showed higher growth than Gxf1 at higher xylose concentrations (Fig. 3b).

1.4.5 Yeast fermentation with chosen xylose transporters

Fermentation assays were done in EBY_Xyl1 for the four transporters in media containing 1% xylose as the sole carbon source, as well as for Gxf1 and empty pRS426 vector (positive and negative controls, respectively). Based on the results shown in Fig. 4, SpX, SpG and SpH conferred superior growth capability compared to the traditional Gxf1 transporter. Cells expressing SuL had a smaller rate of growth. A similar pattern was seen for xylose consumption, where SpH conferred a slightly higher assimilation rate than the other transporters, and SuL being the slowest.

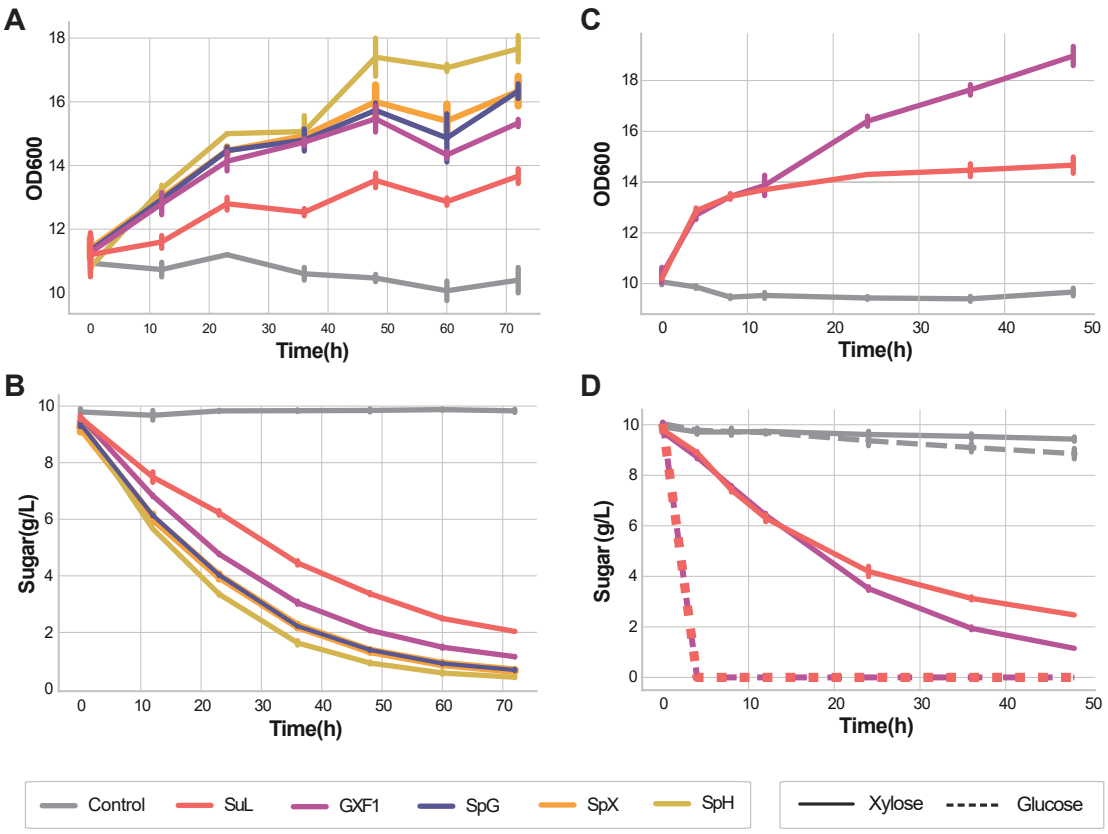


Fig 4. Comparative fermentation assays of EBY_Xyl1 expressing different transporters in xylose (full lines) or glucose (dashed lines). **a** Growth of EBY_Xyl1 during xylose fermentation. **b** Xylose consumption of EBY_Xyl1 cells expressing the transporters over time. Note that SpX does not appear clearly as it overlaps with SpG. **c** Growth of EBY_Xyl1 expressing SuL, GXF1 as positive control and pRS426 (empty vector) as negative control during xylose/glucose co-fermentation and **d** sugar consumption of EBY_Xyl1 expressing SuL, GXF1 as positive control and pRS426 (empty vector) as negative control during xylose/glucose co-fermentation (note that SuL glucose fermentation overlaps with GXF1)

Due to the performance results from the spot-assay in different C6 sugars, simultaneous consumption of xylose and glucose by SuL was evaluated by fermentation of a mixture of 10 g/L each of xylose and glucose., Glucose was entirely consumed on the

first 4 h of experiment, while xylose was slowly consumed during the same period, only increasing after glucose depletion. After 20 h, cells expressing GXF1 also demonstrated more efficiency in transporting xylose than SuL.

1.4.6 Comparative docking of transporters

All four transporters and Gxf1 were modelled through RoseTTAFold in Robetta server [69] for comparative docking using the xylE crystallographic structure bound to xylose or glucose as a comparison basis [25]. Figure 5 and Table 1 outline the docking results when compared to the pose of the ligands on the crystal (lowest root-mean-square deviation of atomic positions—RMSD, the average distance between superimposed structures—obtained between the docked pose and the crystal's ligand pose during self-docking) and their simulated pose. Near identical poses for all transporters were achieved, with RMSDs ranging from 0.6 to 2 Å, which are generally accepted as good modelling outcomes [70]. All sequences had similar affinity to xylose, with SuL having the lowest, SpX, SpH and SpG slightly higher than Gxf1, and xylE having the highest. These results are partially supported by the comparative fermentation in xylose, in which these affinity patterns can be seen on xylose consumption rate and cell growth. All transporters' calculated docking affinity to glucose was higher than to xylose, indicating the typical behaviour of substrate promiscuity and preferential uptake of glucose.

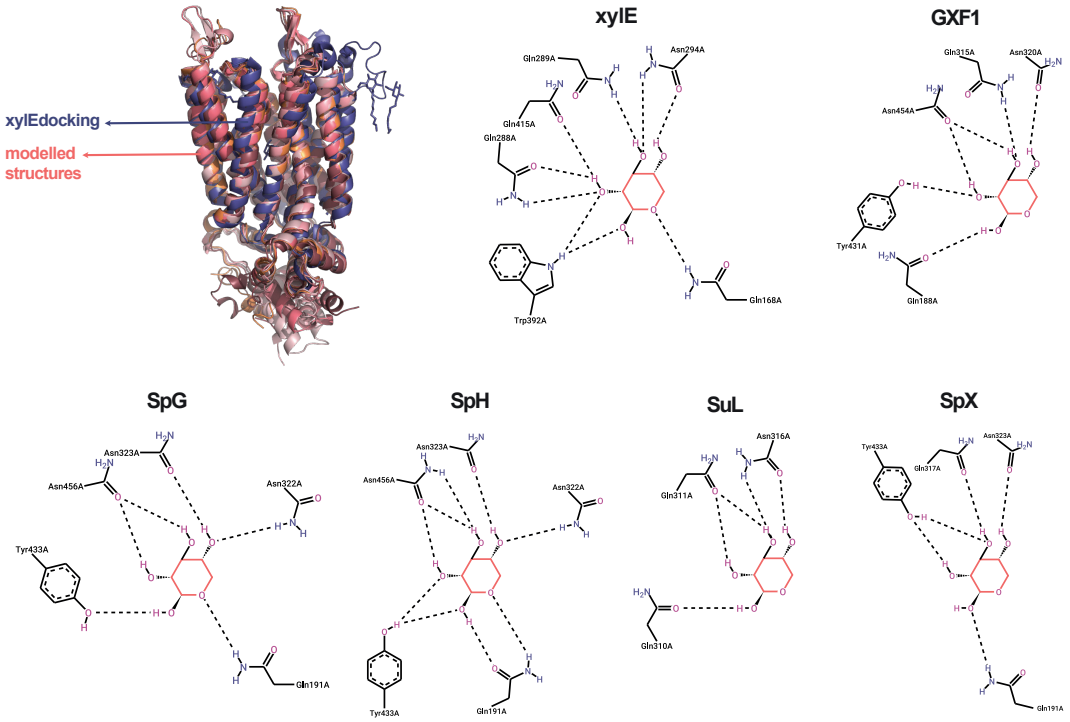


Fig 5. Superimposed structures of xylE coupled with xylose (blue) and predicted structures for the four xylose transporters and GXF1 (pink tones). The 2D representations show the probable interactions between xylose and amino acids in the binding site for each transporter

Table 1. Docking results for the four candidate transporters, xylE (self-docking) and Gxf1

Protein name	**Xylose**		**Glucose**	
	Affinity	Δ RMSD from crystal	Affinity	Δ RMSD from crystal
xylE	-5.8	1.776	-6	0.619
SuL	-5.0	0.948	-5.8	1.166
Gxf1	-5.3	1.085	-5.7	1.296
SpH	-5.4	2.274	-5.5	2.252
SpX	-5.5	2.454	-6	2.668
SpG	-5.5	2.300	-5.6	1.223

Affinity represents the stability of the ligand in the binding site (the more negative the better), and Δ RMSD represents the difference in pose between docked prediction and xylE crystal position

1.5 Discussion

Describing novel transporters is an important step to help unravel the underlying causes in which a sugar transporter is able to transport xylose while another does not show this capacity. The use of computational approaches, such as with machine learning or comparative genomics, have become powerful tools in this search effort. Some algorithms have been proposed to predict different transporter classes based on their function to facilitate classification and categorization [54, 71], however, even though they efficiently categorize membrane transporters, these models aim for a broader classification, which results in not deep enough information regarding function for some specific purposes such as sugar transport capacity discrimination. This work presents a classification model with the purpose of distinguishing sugar transporters in their ability to transport xylose.

The use of oversampling techniques coupled with increasing the prediction threshold were able to create a trustful model which identified 25 potential xylose transporters, from which four were experimentally validated. This shows that, even with a restrictive baseline threshold, many transporters from a diverse group of species were returned, highlighting the potential of different microorganisms, many of them rarely studied with an applied biotechnological view, in supplying candidate genes for bio-industrial applications. Regarding other sequences not chosen for further investigation, some transporters were surprising to appear as positive from our model, such as Pickudriav5544 and Pickudriav5977 from *Pichia kudriavzevii*, which on a first literature screen for the 182 yeast phylogenetics study appears as a species incapable of xylose transport. A second

screening showed that they are able to utilize xylose [72], which increased confidence that the altering of the threshold for the models' predictions effectively removed false positives.

The model also highlighted 13 features as most important for its predictive capability, from which two, PSSM profiling and AAindex, were also used found in previous studies [71]. Interestingly, the model also highlighted the HMM score feature, originally developed in this work. Briefly, this feature was generated by isolating the non-cytoplasmic region of the known xylose transporters used for model creation through sequence alignment followed by comparison with the InterproScan results for *Debaryomyces fabryi* Xylhp (Uniprot accession Q64L87). Another interesting feature was GFV tripeptides, which are located on transmembrane portions of the transporters, but their direct relation to xylose transport is unclear. Nonetheless, all predicted transporters had this tripeptide conformation ranging from 1 to 3 groups. PSSM and the custom HMM features highlight and hint that there is a hidden motif associated with xylose affinity, which due to the nature of the boosting algorithm was not yet humanly interpretable, but with future improvements of the model may come to light. These in silico results are also in accordance with previous experimental works that have also shown that xylose affinity is correlated with sequence alterations, key motifs, and amino acid interactions with the sugar ligand [15, 37].

One interesting pattern that we detected posteriorly to choosing the candidates was that SpX, SpG and SpH were on the same clade on the fam10 phylogeny and also form a monophyletic clade on the species' phylogeny, highlighting an overlap of past adaptations (phylogenomic analysis) and recent patterns (machine learning analysis). This pattern overlap can also be seen on the probable convergent evolution of the site found under positive selection between the 3 *Spathaspora* chosen candidates and the *Sugiyamaella* transporter, as Sugiyamaella diverged much earlier but still has the same adaptations as the *Spathaspora* transporters. Moreover, the HMM feature was created on the non-cytoplasmic domains of the known xylose transporters and the site under positive selection is also on one of these non-cytoplasmic domains, again highlighting an overlap between evolutionary marks and more recent sequence attributes. As mentioned before, this site is located on the N-terminal region of the first transmembrane helix, which may have some function associated with stabilizing the rocker-switch mechanism when the transporter is active. Mutating this residue in future studies could help to understand more its role for sugar transport, as structural studies of MFS transporters, such as xylE, have focused on mutating amino acids associated with the predicted sugar binding sites [25, 73, 74]. While machine learning and comparative genomics have been used before separately to classify or describe transporters, to our knowledge this is the first study that associates both strategies and apply them to a bio-industrial challenge.

Regarding the experimental validation, spot-assay results were surprising for

SuL, as there was almost no growth in C6 sugars glucose, fructose, and mannose, while growth in xylose and galactose was restored. Yeasts expressing the four candidates also showed greater growth compared to Gxf1 on concentrations above 10 g/L of xylose indicating that the chosen candidates could be viable for industrial use, as lignocellulosic biomass contains higher xylose concentrations than the condition where Gxf1 is comparable to the other proteins [75, 76], which translates into a higher xylose concentration during industrial fermentations [77] where these transporters have greater activity. However, future studies using *S. cerevisiae* strains adapted to industrial conditions would be required to further validate these candidates. Fermentation assays were also interesting, as SpX, SpG and SpH were all slightly more efficient than the widely used Gxf1 transporter, with SuL lagging. Co-fermentation assays of SuL and Gxf1 showed that the latter has a slightly superior consumption pattern of xylose, using glucose during the first 4 h, and only then using xylose, as expected. Gxf1 also conferred higher growth in the C5 sugar. Again however, an evaluation on an industrial strain with higher xylose and glucose concentrations on the media would reveal the industrial potential of these candidates, as coupled with spot-assay results these patterns suggest that these four xylose transporters can expand the repertoire for industrial use and build a strong case for the model's use on prospecting novel candidates. Also, even though SuL showed a lower consumption rate both in xylose and during co-fermentation when compared to Gxf1, the lack of growth in glucose and fructose as seen in the spot-assay is a rare phenotype for sugar transporters and could indicate an interesting target for mutagenesis or directed evolution aiming to increase its xylose consumption rate. By using a combination of sugar transporters with different affinities, a future industrial yeast strain could metabolize C6 and C5 sugars more effectively, increasing the viability of the 2G process.

As an attempt to consolidate all results obtained in this study, the 3D structure of the experimentally evaluated transporters was created, with docking analysis coupled with glucose and xylose. Docking analysis successfully modelled both xylose and glucose poses for the evaluated transporters, which gave us a bigger confidence on the affinity calculations. These affinity results were confirmed during fermentation, where slight growth differences were seen in accordance with the slight differences in predicted affinity.

Finally, with future advances on describing novel xylose-transporting proteins and the increase of sequenced genomes, the model can be improved and become an important tool for researchers on helping to prospect industrial candidate transporters.

1.6 Conclusions

2G ethanol is a promising energy matrix alternative for current and future needs. One challenge for this technology's viability is an efficient and uninhibited transport

of pentose sugars into yeast cells, which drives the search for novel and capable xylose transporters to be engineered into industrial S. cerevisiae, the main organism used for this kind of fermentation. The coupled machine learning and comparative genomics approach presented here yielded several xylose transporter candidates, from which four were experimentally tested against a wide range of sugars. The dimensionality reduction by feature selection highlighted that the most important features were related to HMM and PSSM profiles, indicating that xylose transport can in part be explained by amino acid patterns in the non-cytoplasmic domains of the proteins, especially the pore and binding sites, a result also seen in previous point mutation studies and descriptions of known transporters' structures, indicating the conformity of the model with previous studies in literature. All transporters tested successfully transported xylose, most of them in rates superior to the traditional Gxf1, one of the best-known heterologous xylose transporters in literature, and all conferred higher cell growth at larger xylose concentrations. Docking analysis showed a similar pattern, with SuL having the lowest affinity to xylose and the other three transporters having a slightly higher affinity than Gxf1.

For future studies, the model's predictive capability should be provided with data arising from new xylose transporter characterizations, as well as attempts to create models by adding information of known xylose transporters from other organisms, such as Arabidopsis thaliana. We believe that not only researchers interested in prospecting novel xylose-transporting candidates for industrial application can already make use of the model to aid their selection of best targets for wet-lab evaluation, but also understanding xylose transport on a broader scale then fungi and bacteria, with the help of this model, it will be possible to better understand and reveal the intricacies of xylose transport.

1.7 Methods

1.7.1 182 genomes dataset

Genomes were retrieved from NCBI, based on if they were the representative genome for that species. 30 genomes had no coding sequences prediction, and so had their genes predicted by using AUGUSTUS 3.3.2 [78] and GeneMark-ES Suite 4.32 [79] separately and reconciled with Evidence Modeler 1.1.1 [72]. Genes were also filtered by their longest ORF via Transdecoder and by having at least 80 amino acids in the sequence. Genome completeness and success of gene prediction was analysed by utilizing the BUSCO v3 [80] Saccharomycotina dataset, which comprised conserved genes from this group and must be found on the data for a successful gene prediction. Additional file 6: fig. S1 shows the results for the BUSCO analysis.

1.7.2 Phylogenomic analysis

Genes were clustered into gene families by means of Orthofinder 2.2 [55]. Transporter families were retrieved through the HMM MFS_1 and MFS_5 profiles from Pfam [64], and through known xylose transporters (XUT1, GXF1, GXS1, HXT7, Xylhp, XUT3, xylE, Cs3894, Cs4130) as baits for a BLAST search. Multi-sequence alignments were undertaken with MAFFT [81] for protein sequences (L-INS-i), and with MACSE 2.01 [82] by anchoring with the protein alignment for CDS. Alignments were trimmed using Trimal 1.4.1 [83] for phylogenetic inferences of the conserved domains, as due to the nature of the dataset (too many sequences from heterogenous groups) there were many gaps. Phylogenetic inferences were done through Maximum Likelihood with IQTree 1.6.12 [84] running 1000 bootstraps. Selection analysis was done by marking sequences that were output from the Machine Learning model as foreground and running HYPHY MEME 2.0.1 [85].

1.7.3 Machine Learning

Machine learning modelling usually undergoes the following steps: data clean-up and division into training and testing datasets, feature extraction and selection, model training, and evaluation. After clean-up, all intermediate steps are done on the training dataset and evaluation is done with the testing dataset. This separation of training and testing allows for a faithful evaluation of a model's metrics by isolating some of the data in such a way that testing is done on part of the dataset upon which the model has no bias. Also, all these steps can be done using different machine learning algorithms and it is recommended to test several models using different algorithmical approaches and selecting the best performant. Feature extraction in the case of protein modelling represents decomposing the amino acid sequence into different descriptors that either mathematically explain the sequence or highlight some trait of interest, while feature selection is used for dimensionality reduction, computational optimization and highlighting the importance of certain features for classification.

Sugar transporters from fungi and bacteria were retrieved from Uniprot and TCDB databases. CD-HIT [86] was done to remove proteins with more than 80% sequence similarity, except for the known xylose transporters (experimentally validated by other studies), which were manually re-added to the dataset if removed. Xylose transporter sequences with their respective publication are shown in Additional file 2: Table S2. Most features were extracted with the protr package [87], which generates many numerical explainers of a given protein sequence. Also, an HMM feature was calculated by aligning the xylose transporters and using the sequence of Xylhp from *Debaryomyces fabryi* (Uniprot accession Q64L87) to predict domains and important sites through Interproscan; then, the non-cytoplasmic domains and probable sugar binding sites were isolated from the

alignment and the HMM profile was created. We assumed as all sequences are aligned, the binding sites would be roughly in the same position. Other features added were the PS0021 and PS00217 sugar transport signatures from PROSITE database [88] using ScanProsite [89], the protein existence evidence, which sugars it transports, protein annotation, and if there is evidence in literature for xylose transport.

Following feature extraction and clean-up, sequences were divided into training and testing datasets through scikit-learn's 0.21.2 [58] train_test_split with 70% used for training and 30% for testing.

Feature selection was done by Recursive Feature Elimination with Cross-Validation (RFECV) by using a Gradient Boosting Decision Tree classifier, implemented by XGBoost 0.82 [90]. Feature importance visualization was done using Yellowbrick 0.9.1 [91] or SHAP 0.29.3 explainers [92].

Some statistical transformations using oversampling were attempted to mitigate dataset imbalance, at the cost of some overfitting of the data. Additional file 7: figure S2 shows UMAP 0.3.9 [93] spatial distribution of samples after oversampling through Random Oversampling, SMOTE, SMOTEEEN, SMOTETomek and ADASYN. Except for Random Oversampling, all these transformers use a nearest neighbour approach to add a synthetic new sample to the data, which is related to the parameters of its neighbors. A model from all these attempts was made, however only SMOTEEEN was taken further as the evaluated model metrics were more satisfactory.

Model evaluation was done through usual metrics, such as accuracy, precision-recall, AUC, ROC curve, Balanced Accuracy and MCC, however, we were also attentive to brute numbers, because of the dataset imbalance distorting metric results. False positives were penalized by increasing the classification threshold of the xylose-positive class to 0.98, and this restrictive model was used for choosing candidates. The four transporter families from our phylogenomics dataset had the 13 most important features calculated and the model was ran. Sequences were chosen based on our knowledge if their respective species is a known fermenter or consumer of xylose.

All code used for model creation and data engineering can be found at https://gitlab.com/Matt_BF/Xylose_Transporter_ML.

1.7.4 Strains and constructions

Strain EBY_Xyl1 was constructed from EBY.VW4000 by inserting an expression cassette containing the genes XYL1 and XYL2 from *S. stipitis* and an additional copy of xylulokinase (XKS1) under control of different promoters of the glycolytic pathway of *S. cerevisiae* as previously described [47]. Synthesized SpG, SpH, SpX and SuL were cloned into pRS426 at the EcoRI and NotI sites flanked with the promoter and terminator

regions from THD1 gene and further transformed into EBY_Xyl1 through the LiAc/SS-DNA/PEG protocol [94]. Transformants were selected in YNB medium lacking uracil. The transformation was confirmed by PCR using primers for the coding sequence of each gene (Additional file 3: Table S3).

1.7.5 Media and culture conditions

Yeast cells were grown on liquid YP medium (10 g/L yeast extract and 20 g/L peptone) supplemented with 20 g/L D-glucose (YPD) for cell propagation or 20 g/L D-xylose (YPX) for xylose growth analysis. Transformed cells were grown at 30 °C in complete synthetic media YNB (6.7 g/L yeast nitrogen base without amino acids, Difco) supplemented with 1 g/L drop-out without uracil, 20 g/L glucose and 20 g/L agar [75]. YP was autoclaved at 121 °C for 20 min and YNB was filter-sterilized using 0.2-μm bottle-top filters. Strain EBY.VW4000, kindly supplied by Prof. Eckhard Boles from Goethe university [39], and strain EBY_Xyl1 were grown in YNB with D-maltose instead of D-glucose.

1.7.6 Fermentations

Yeast strains were pre-grown on YNB supplemented with 5 g/L of casamino acids (Difco), 1 g/L of tryptophan (Sigma) and 50 g/L of d-maltose for 24 h. Cells were then harvested by centrifugation, washed three times with sterile water and resuspended to an OD600 of 10. Fermentation experiments were performed aerobically in 250 mL Erlenmeyer flasks using 70 mL of YNB supplemented with 5 g/L of casamino acids, 1 g/L of tryptophan (Sigma) and 10 g/L of xylose. For simultaneous glucose and xylose co-fermentation, 10 g/L of both sugars were used. The cells were incubated at 30 °C/200 rpm. Experiments were performed in triplicate and samples were collected to measure optical density and for HPLC analysis.

1.7.7 Molecular Docking

Molecular docking analysis was done using Autodock-Vina 1.1.2, ran via UCSF Chimera 1.15. Transporter structures for Gxf1, SuL, SpG, SpH and SpX were modelled through ROSETTAFold via the Robetta server [69], with the lowest angstrom error estimate models chosen for docking, and the glucose and xylose ligands were obtained from PubChem (IDs 5793 and 135191, respectively). The xylE crystal structure bounded to xylose (PDB code 4GBY) or glucose (PDB code 4GBZ) was used as the reference for self-docking and for interpretation of the other transporters (evaluation and comparison of ligand position on the candidate transporters and during self-docking, as in the closest the ligand poses during docking to the pose from the xylE crystal the better). Ten docking runs were done for each transporter and the one with the lowest RMSD from the xylE

crystal was chosen. Comparison of ligand position and poses was done with DockRMSD [95]. Visualization of docking results and ligand positions was done with pyMOL, and the 2D ligand interactions were extracted on the Protein-Plus web server [96].

1.8 References

1. Zaldivar J, Nielsen J, Olsson L. Fuel ethanol production from lignocellulose: a challenge for metabolic engineering and process integration. Appl Microbiol Biotechnol. 2001;56:17–34.

2. Gírio FM, Fonseca C, Carvalheiro F, Duarte LC, Marques S, Bogel-Łukasik R. Hemicelluloses for fuel ethanol: a review. Biores Technol. 2010;101:4775– 800. https://doi.org/10.1016/j.biortech.2010.01.088.

3. Dias MOS, Junqueira TL, Cavalett O, Pavanello LG, Cunha MP, Jesus CDF, et al. Biorefneries for the production of frst and second generation ethanol and electricity from sugarcane. App Energy. 2013;109:72–8. https://doi.org/10.1016/j.apenergy.2013.03.081.

4. Balat M. Production of bioethanol from lignocellulosic materials via the biochemical pathway : a review. Energy Convers Manag. 2011;52:858–75. https://doi.org/10.1016/j.enconman.2010.08.013.

5. Zhao Z, Xian M, Liu M, Zhao G. Biochemical routes for uptake and conversion of xylose by microorganisms. Biotechnol Biofuels. 2020. https://doi. org/10.1186/s13068-020-1662-x.

6. Maga D, Thonemann N, Hiebel M, Sebastião D, Lopes TF, Fonseca C, et al. Comparative life cycle assessment of frst- and second-generation ethanol from sugarcane in Brazil. Int J Life Cycle Assess. 2019;24:266–80.

7. Gírio FMM, Fonseca C, Carvalheiro F, Duarte LCC, Marques S, BogelŁukasik R. 2010 Hemicelluloses for fuel ethanol: a review. Biores Technol. 2010;101:4775–800. https://doi.org/10.1016/j.biortech.2010.01.088.

8. dos Santos LV, de Barros Grassi MC, Gallardo JCM, Pirolla RAS, Calderón LL, de Carvalho-Netto OV, et al. Second-generation ethanol: the need is becoming a reality. Ind Biotechnol. 2016;12:40–57. https://doi.org/10. 1089/ind.2015.0017.

9. Jefries TW. Engineering yeasts for xylose metabolism. Curr Opin Biotechnol. 2006;17:320–6.

10. Botstein D, Fink GR. Yeast: an experimental organism for 21st century biology. Genetics. 2011;189:695–704. https://doi.org/10.1534/genetics.111. 130765.

11. Cunha JT, Soares PO, Romaní A, Thevelein JM, Domingues L. 2019 Xylose fermentation efciency of industrial Saccharomyces cerevisiae yeast with separate or com-

bined xylose reductase/xylitol dehydrogenase and xylose isomerase pathways. Biotechnol Biofuels. 2019;12:1–14. https:// doi.org/10.1186/s13068-019-1360-8.

12. Hua Y, Wang J, Zhu Y, Zhang B, Kong X, Li W, et al. Release of glucose repression on xylose utilization in Kluyveromyces marxianus to enhance glucose-xylose co-utilization and xylitol production from corncob hydrolysate. Microb Cell Fact. 2019. https://doi.org/10.1186/s12934-019-1068-2.

13. Lane S, Xu H, Oh EJ, Kim H, Lesmana A, Jeong D, et al. Glucose repression can be alleviated by reducing glucose phosphorylation rate in Saccharomyces cerevisiae. Sci Rep. 2018;8(1):2613.

14. Brink DP, Borgström C, Persson VC, Osiro KO, Gorwa-Grauslund MF. D-xylose sensing in Saccharomyces cerevisiae: Insights from D-glucose signaling and native D-xylose utilizers [Internet]. Int J Mol Sci. 2021;22(22):12410.

15. Farwick A, Bruder S, Schadeweg V, Oreb M, Boles E. Engineering of yeast hexose transporters to transport D-xylose without inhibition by D-glucose. Proc Natl Acad Sci. 2014;111:5159–64.

16. Wieczorke R, Krampe S, Weierstall T, Freidel K, Hollenberg CP, Boles E. Concurrent knock-out of at least 20 transporter genes is required to block uptake of hexoses in Saccharomyces cerevisiae. FEBS Lett. 1999;464:123–8.

17. Hamacher T, Becker J, Gárdonyi M, Hahn-Hägerdal B, Boles E. Characterization of the xylose-transporting properties of yeast hexose transporters and their infuence on xylose utilization. Microbiology. 2002;148:2783–8. https://doi.org/10.1099/00221287-148-9-2783.

18. Weierstall T, Hollenberg CP, Boles E. Cloning and characterization of three genes (SUT1–3) encoding glucose transporters of the yeast Pichia stipitis. Mol Microbiol. 1999;31:871–83. https://doi.org/10.1046/j.1365-2958.1999. 01224.x.

19. Leandro MJ, Gonçalves P, Spencer-Martins I. Two glucose/xylose transporter genes from the yeast Candida intermedia : frst molecular characterization of a yeast xylose–H + symporter. Biochem J. 2006;395:543–9.

20. Ferreira D, Nobre A, Silva ML, Faria-Oliveira F, Tulha J, Ferreira C, et al. XYLH encodes a xylose/H+ symporter from the highly related yeast species Debaryomyces fabryi and Debaryomyces hansenii. FEMS Yeast Res. 2013;13:585–96.

21. Reifenberger E, Boles E, Ciriacy M. Kinetic characterization of individual hexose transporters of Saccharomyces cerevisiae and their relation to the triggering mechanisms of glucose repression. Eur J Biochem. 1997;245:324–33. https://doi.org/10.1111/j.1432-1033.1997.00324.x.

22. Diderich JA, Schepper M, van Hoek P, Luttik MAH, van Dijken JP, Pronk

JT, et al. Glucose uptake kinetics and transcription of HXTGenes in chemostat cultures of Saccharomyces cerevisiae. J Biol Chem. 1999;274:15350–9.

23. Young E, Poucher A, Comer A, Bailey A, Alper H. Functional survey for heterologous sugar transport proteins, using Saccharomyces cerevisiae as a host. App Environ Microbiol. 2011;77(10):3311.

24. Maier A, Völker B, Boles E, Fuhrmann GF. Characterisation of glucose transport in Saccharomyces cerevisiae with plasma membrane vesicles (countertransport) and intact cells (initial uptake) with single Hxt1 Hxt2 Hxt3 Hxt4 Hxt6 Hxt7 or Gal2 transporters. FEMS Yeast Res. 2002;2:539–50.

25. Sun L, Zeng X, Yan C, Sun X, Gong X, Rao Y, et al. Crystal structure of a bacterial homologue of glucose transporters GLUT1–4. Nature. 2012;490:361–6.

26. Wambo TO, Chen LY, Phelix C, Perry G. Afnity and path of binding xylopyranose unto E. coli xylose permease. Biochem Biophys Res Commun. 2017;494:202–6.

27. Madej MG, Sun L, Yan N, Kaback HR. Functional architecture of MFS D-glucose transporters. Proc Natl Acad Sci. 2014. https://doi.org/10.1073/ pnas.1400336111.

28. Shin HY, Nijland JG, de Waal PP, de Jong RM, Klaassen P, Driessen AJM. An engineered cryptic Hxt11 sugar transporter facilitates glucose-xylose co-consumption in Saccharomyces cerevisiae. Biotechnol Biofuels. 2015. https://doi.org/10.1186/s13068-015-0360-6.

29. Subtil T, Boles E. Competition between pentoses and glucose during uptake and catabolism in recombinant Saccharomyces cerevisiae. Biotechnol Biofuels. 2012;5:14. https://doi.org/10.1186/1754-6834-5-14.

30. Donzella L, Varela JA, Sousa MJ, Morrissey JP. Identifcation of novel pentose transporters in Kluyveromyces marxianus using a new screening platform. FEMS Yeast Res. 2021;21(4):26.

31. Hector RE, Qureshi N, Hughes SR, Cotta MA. Expression of a heterologous xylose transporter in a Saccharomyces cerevisiae strain engineered to utilize xylose improves aerobic xylose consumption. Appl Microbiol Biotechnol. 2008;80:675–84.

32. de Sales BB, Scheid B, Gonçalves DL, Knychala MM, Matsushika A, Bon EPS, et al. Cloning novel sugar transporters from Schefersomyces (Pichia) stipitis allowing d-xylose fermentation by recombinant Saccharomyces cerevisiae. Biotechnol L. 2015;37:1973–82. https://doi.org/10.1007/ s10529-015-1893-2.

33. dos Reis TF, de Lima PBA, Parachin NS, Mingossi FB, de Castro Oliveira JV, Ries LNA, et al. Identifcation and characterization of putative xylose and cellobiose transporters in Aspergillus nidulans. Biotechnol Biofuels BioMed Cent. 2016;9:1–19.

34. Lane S, Xu H, Oh EJ, Kim H, Lesmana A, Jeong D, et al. Glucose repression can be alleviated by reducing glucose phosphorylation rate in Saccharomyces cerevisiae. Sci Rep. 2018. https://doi.org/10.1038/ s41598-018-20804-4.

35. Caballero A, Ramos JL. Enhancing ethanol yields through D-xylose and L-arabinose co-fermentation after construction of a novel high efcient L-arabinose-fermenting Saccharomyces cerevisiae strain. Microbiology. 2017;163:442–52.

36. Li H, Schmitz O, Alper HS. Enabling glucose/xylose co-transport in yeast through the directed evolution of a sugar transporter. App Microbiol Biotechnol. 2016;100:10215–23. https://doi.org/10.1007/s00253-016-7879-8.

37. Wang M, Yu C, Zhao H. Directed evolution of xylose specifc transporters to facilitate glucose-xylose co-utilization. Biotechnol Bioeng. 2016;113:484–91. https://doi.org/10.1002/bit.25724.

38. Nijland JG, Shin HY, de Jong RM, de Waal PP, Klaassen P, Driessen AJM. Engineering of an endogenous hexose transporter into a specifc D-xylose transporter facilitates glucose-xylose co-consumption in Saccharomyces cerevisiae. Biotechnol Biofuels. 2014;7:168. https://doi.org/10. 1186/s13068-014-0168-9.

39. Kuanyshev N, Deewan A, Jagtap SS, Liu J, Selvam B, Chen LQ, et al. Identifcation and analysis of sugar transporters capable of co-transporting glucose and xylose simultaneously. Biotechnol J. 2021. https://doi.org/10. 1002/biot.202100238.

40. Young EM, Comer AD, Huang H, Alper HS. A molecular transporter engineering approach to improving xylose catabolism in Saccharomyces cerevisiae. Metabol Eng. 2012. https://doi.org/10.1016/j.ymben.2012.03. 004.

41. Wijsman M, Marques WL, Hettinga JK, van den Broek M, de la CortésCort-Torre P, Mans R, et al. A toolkit for rapid CRISPR-SpCas9 assisted construction of hexose-transport-defcient Saccharomyces cerevisiae strains. FEMS Yeast Res. 2019;19(1):107.

42. Reider Apel A, Ouellet M, Szmidt-Middleton H, Keasling JD, Mukhopadhyay A. Evolved hexose transporter enhances xylose uptake and glucose/ xylose co-utilization in Saccharomyces cerevisiae. Sci Rep. 2016;6(1):19512.

43. Lin Z, Li WH. Expansion of hexose transporter genes was associated with the evolution of aerobic fermentation in yeasts. Mol Biol Evol. 2011;28:131–42.

44. Wohlbach DJ, Kuo A, Sato TK, Potts KM, Salamov AA, Labutti KM, et al. Comparative genomics of xylose-fermenting fungi for enhanced biofuel production. Proc Natl Acad Sci USA. 2011;108:13212–7.

45. Riley R, Haridas S, Wolfe KH, Lopes MR, Hittinger CT, Göker M, et al. Comparative genomics of biotechnologically important yeasts. Proc Natl Acad Sci. 2016;113:9882–7. https://doi.org/10.1073/pnas.1603941113.

46. Borelli G, Fiamenghi MB, dos Santos LV, Carazzolle MF, Pereira GAG, José J, et al. Positive selection evidence in xylose-related genes suggests methylglyoxal reductase as a target for the improvement of yeasts' fermentation in industry. Genome Biol Evolution. 2019;11:1923–38. https://doi.org/ 10.1093/gbe/evz036.

47. Bueno JGR, Borelli G, Corrêa TLR, Fiamenghi MB, José J, de Carvalho M, et al. Novel xylose transporter Cs4130 expands the sugar uptake repertoire in recombinant Saccharomyces cerevisiae strains at high xylose concentrations. Biotechnol Biofuels. 2020;13:145.

48. Lin HH, Han LY, Cai CZ, Ji ZL, Chen YZ. Prediction of transporter family from protein sequence by support vector machine approach. Proteins. 2006. https://doi.org/10.1002/prot.20605.

49. Li H, Dai X, Zhao X. A nearest neighbor approach for automated transporter prediction and categorization from protein sequences. Bioinformatics. 2008;24:1129–36.

50. Bhasin M, Raghava GPS. Classifcation of nuclear receptors based on amino acid composition and dipeptide composition. J Biol Chem. 2004. https://doi.org/10.1074/jbc.M401932200.

51. Chou KC, Cai YD. Using functional domain composition and support vector machines for prediction of protein subcellular location. J Biol Chem. 2002. https://doi.org/10.1074/jbc.M204161200.

52. Sarda D, Chua GH, Li K, bin, Krishnan A. pSLIP: SVM based protein subcellular localization prediction using multiple physicochemical properties. Bioinformatics. 2005. https://doi.org/10.1186/1471-2105-6-152.

53. Lv Z, Jin S, Ding H, Zou Q. A random forest sub-golgi protein classifer optimized via dipeptide and amino acid composition features frontiers in bioengineering and biotechnology. Frontiers. 2019;0:215.

54. Gromiha MM, Yabuki Y. Functional discrimination of membrane proteins using machine learning techniques. Bioinformatics. 2008;9:1–8.

55. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 2015. https://doi.org/10.1186/s13059-015-0721-2.

56. Bateman A, Martin MJ, O'Donovan C, Magrane M, Alpi E, Antunes R, et al. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2017;45:D158–69.

57. Saier MH, Tran CV, Barabote RD. TCDB: the transporter classifcation database for membrane transport protein analyses and information. Nucleic Acids Res. 2006;34:D181–6.

58. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. J Mach Learn Res. 2012;12:2825–30.

59. Gromiha MM. Protein sequence analysis. protein. Bioinformatics. 2010;0:29–62.

60. Dayhof MO, Schwartz RM. Chapter 22: a model of evolutionary change in proteins. In: Atlas of protein sequence and structure. Washington: National Biomedical Research Foundation; 1978.

61. Bigelow CC. On the average hydrophobicity of proteins and the relation between it and protein structure. J Theor Biol. 1967;16:187–211.

62. van Westen GJP, Swier RF, Cortes-Ciriano I, Wegner JK, Overington JP, Jzerman API, et al. Benchmarking of protein descriptor sets in proteochemometric modeling (part 2): modeling performance of 13 amino acid descriptor sets. J Cheminform. 2013. https://doi.org/10.1186/ 1758-2946-5-42.

63. Chou KC. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics. 2005. https://doi.org/10.1093/ bioinformatics/bth466.

64. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: The protein families database in 2001. Nucleic Acids Res. 2021. https://doi.org/10.1093/nar/gkaa913.

65. Bellasio M, Peymann A, Steiger MG, Valli M, Sipiczki M, Sauer M, et al. Complete genome sequence and transcriptome regulation of the pentose utilizing yeast Sugiyamaella lignohabitans. FEMS Yeast Res. 2016. https://doi.org/10.1093/femsyr/fow037.

66. Trichez D, Steindorf AS, Soares CEVF, Formighieri EF, Almeida JRM. Physiological and comparative genomic analysis of new isolated yeasts Spathaspora sp JA1 and Meyerozyma caribbica JA9 reveal insights into xylitol production. FEMS Yeast Res. 2019. https://doi.org/10.1093/femsyr/ foz034.

67. Mixao V, Hegedusova E, Saus E, Pryszcz LP, Cillingova A, Nosek J, et al. Genome analysis of Candida subhashii reveals its hybrid nature and dual mitochondrial genome conformations. DNA Res. 2021. https://doi.org/10. 1093/dnares/dsab006.

68. Runquist D, Fonseca C, Rådström P, Spencer-Martins I, Hahn-Hägerdal B. Expression of the Gxf1 transporter from Candida intermedia improves fermentation performance in recombinant xylose-utilizing Saccharomyces cerevisiae. Appl Microbiol Biotechnol. 2009. https://doi.org/10.1007/ s00253-008-1773-y.

69. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, et al. Accurate prediction of protein structures and interactions using a threetrack neural network. Science. 1979;2021(373):871–6.

70. Ramírez D, Caballero J. Is It reliable to take the molecular docking top scoring position as the best solution without considering available structural data? Molecules. 2018;23:1038.

71. Mishra NK, Chang J, Zhao PX. Prediction of membrane transport proteins and their substrate specifcities using primary sequence information. PLoS ONE. 2014;9:3–6.

72. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic gene structure annotation using evidencemodeler and the program to assemble spliced alignments. Genome Biol. 2008. https://doi. org/10.1186/gb-2008-9-1-r7.

73. Drew D, North RA, Nagarathinam K, Tanabe M. Structures and general transport mechanisms by the major facilitator superfamily (MFS). Chem Rev. 2021. https://doi.org/10.1021/acs.chemrev.0c00983.

74. Wisedchaisri G, Park M-S, Iadanza MG, Zheng H, Gonen T. Proton-coupled sugar transport in the prototypical major facilitator superfamily protein XylE. Nat Commun. 2014. https://doi.org/10.1038/ncomms5521.

75. Blomqvist J, South E, Tiukova L, Momeni MH, Hansson H, Ståhlberg J, et al. Fermentation of lignocellulosic hydrolysate by the alternative industrial ethanol yeast dekkera bruxellensis. Lett Appl Microbiol. 2011. https://doi. org/10.1111/j.1472-765X.2011.03067.x.

76. Senatham S, Chamduang T, Kaewchingduang Y, Thammasittirong A, Srisodsuk M, Elliston A, et al. Enhanced xylose fermentation and hydrolysate inhibitor tolerance of schefersomyces shehatae for efcient ethanol production from non-detoxifed lignocellulosic hydrolysate. Springerplus. 2016. https://doi.org/10.1186/s40064-016-2713-4.

77. Carvalho LM, Carvalho-Netto OV, Calderón LL, Gutierrez M, de Assis MA, Mofatto LS, et al. Understanding the diferences in 2G ethanol fermentative scales through omics data integration. FEMS Yeast Res. 2021;21:1–13.

78. Keller O, Kollmar M, Stanke M, Waack S. A novel hybrid gene prediction method employing protein multiple sequence alignments. Bioinformatics. 2011;27:757–63.

79. Besemer J, Lomsadze A, Borodovsky M. GeneMarkS a self-training method for prediction of gene starts in microbial genomes implications for fnding sequence motifs in regulatory regions. Nucl Acids Res. 2001. https://doi.org/10.1093/nar/29.12.2607.

80. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015. https://doi.org/10.1093/bioin formatics/btv351.

81. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772–80.

82. Ranwez V, Harispe S, Delsuc F, Douzery EJP. MACSE: multiple alignment of coding sequences accounting for frameshifts and stop codons Murphy WJ, editor. PLoS ONE. 2011. https://doi.org/10.1371/journal.pone.00225 94.

83. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. TrimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25:1972–3.

84. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: new models and efcient methods for phylogenetic inference in the genomic Era. Mol Biol Evol. 2020;37:1530–4.

85. Murrell B, Wertheim JO, Moola S, Weighill T, Schefer K. Detecting individual sites subject to episodic diversifying selection. PLoS Genet. 2012. https://doi.org/10.1371/journal.pgen.1002764.

86. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinform Oxf Acad. 2006;22:1658–9.

87. Xiao N, Cao DS, Zhu MF, Xu QS. Protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. Bioinformatics. 2015;31:1857–9.

88. Sigrist CJA, de Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, et al. New and continuing developments at PROSITE. Nucl Acids Res. 2013. https:// doi.org/10.1093/nar/gks1067.

89. de Castro E, Sigrist CJA, Gattiker A, Bulliard V, Langendijk-Genevaux PS, Gasteiger E, et al. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucl Acids Res. 2006. https://doi.org/10.1093/nar/gkl124.

90. Chen T, Guestrin C. 2016 XGBoost: a scalable tree boosting system. https://doi.org/10.1145/2939672.2939785.

91. Bengfort B, Bilbro R, Danielsen N, Gray L, McIntyre K, Roman P, et al. 2018 Yellowbrick v0.9 https://zenodo.org/record/1488364.

92. Lundberg SM, Lee SI. A unifed approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30:4766–75.

93. Mcinnes L, Healy J, Melville J. UMAP: Uniform manifold approximation and projection for dimension reduction. J Open Sour Softw. 2018. https:// doi.org/10.21105/joss.00861.

94. Gietz RD. Yeast Transformation by the LiAc/SS Carrier DNA/PEG Method. New York: Humana Press; 2014. https://doi.org/10.1007/ 978-1-4939-1363-3_1.

95. Bell EW, Zhang Y. DockRMSD: an open-source tool for atom mapping and RMSD calculation of symmetric molecules through graph isomorphism. J Cheminform. 2019. https://doi.org/10.1186/s13321-019-0362-7. 96. Schöning-Stierand K, Diedrich K, Fährrolfes R, Flachsenberg F, Meyder A, Nittinger E, et al. ProteinsPlus: interactive analysis of protein–ligand binding interfaces. Nucleic Acids Res. 2020. https://doi.org/10.1093/nar/gkaa2 35.

2 Comparative Genomics of Firmicutes reveals probable adaptations for xylose fermentation in *Thermoanaerobacterium saccharolyticum*

2.1 Abstract

Second-generation (2G) ethanol is one potential biofuel that could be used to achieve the goal of reducing greenhouse gas emissions. Many challenges still need to be overcome for the feasibility of this technology, most of them related to consumption of xylose, a pentose sugar not easily metabolized by industrial microorganisms. Thus, exploring genes, pathways and other organisms that can ferment xylose is a strategy implemented to solve industrial bottlenecks. *Thermoanaerobacterium saccharolyticum* (*T. sac*) is an organism from the firmicutes phylum, capable of naturally fermenting compounds of industrial interest, such as xylan and xylose. Understanding evolutionary adaptations may help not only to solidify this bacterium as a potential substitute to the yeast *Saccharomyces cerevisiae* in industry, but also bring novel genes and information that can be used for yeast, enhance its fermenting capabilities, and increase production of current bio-platforms. This study presents a deep evolutionary study of members of the firmicutes clade, focusing on adaptations that may be related to overall fermentation metabolism, especially for xylose fermentation. One highlight is the finding of positive selection on a xylose binding protein of the xylFGH operon, close to the annotated sugar binding site, with this protein already being found to be expressed in xylose fermenting conditions in a previous study. Results from this study can serve as basis for searching for candidate genes to use in industrial strains or to improve *T. sac* as a new microbial cell factory, which may help to solve current problems found in the biofuels industry.

2.2 Introduction

The need to restructure the global energy matrix and mitigate greenhouse gas emissions has led efforts to find clean alternatives to fossil fuels. One approach is the use of microorganisms for fermentation of lignocellulosic biomasses such as sugarcane bagasse and straw to ethanol, which is particularly interesting as it helps to alleviate competition with foods production. This strategy is known as second generation (2G) ethanol production.

The main challenges associated with 2G ethanol are the forming of fermentation inhibitors, such as HMF and acetate, after the feedstock pretreatment step for exposure of its sugars, finding optimal enzymes for breaking sugars into usable monomers, and lack of proper consumption of pentose sugars, such as xylose, contained on these feedstocks by organisms used industrially, such as the yeast *Saccharomyces cerevisiae*.

Xylose metabolism mainly follows two pathways: an oxireductive pathway, comprising a conversion of xylose into xylitol by the enzyme xylose reductase, followed by a conversion of xylitol into xylulose by xylitol dehydrogenase, and finally xylulokinase converts xylulose into xylulose-5-P, which enters the pentose phosphate pathway, ending up in the glucose pathway. In many organisms, including industrial yeast, this pathway has some issues regarding cofactor imbalance on the first two steps, which causes accumulation of xylitol [40]. The other pathway is similar to the first but comprises a single step between xylose and xylulose, done by xylose isomerase, and is predominantly found in bacteria [83]. One bottleneck common to both pathways is related to pentose transport, in which sugar transporters preferentially uptake glucose in detriment of xylose, turning the 2G process unfeasible due to fermentation time increase.

Thermoanaerobacterium saccharolyticum (hereafter called *T. sac*) is a bacterium from the firmicutes phylum, found originally in hot springs around Yellowstone National Park [41], capable of naturally fermenting compounds of industrial interest, such as xylan and xylose, and engineered to produce ethanol at higher yields [63]. This organism, among other thermophilic bacteria, has gained some attention over the last decade as an alternative for industrial fermentation, as it grows in higher temperature, similar to those used for pre-treatment enzymes, and can co-ferment both cellulose and hemicellulose sugars [15], a trait seen in many *Thermoanaerobacter* [43,72], which means potential for simultaneous saccharification and fermentation. Also, its ease of transformation confers an additional advantage for metabolic engineering [46,47]. Studies have explored the underlying mechanisms of pentose metabolism in *T. sac*, such as inactivating redox sensing molecules to alleviate alcohol dehydrogenase repression [84], deleting genes that create undesirable byproducts such as acetate and lactic acid [63], discovering essential genes [14], and elucidating fermentation products [32]. However, an evolutionary approach to understand key adaptations to fermentation stresses and describing candidate genes has not been reported for this organism. Evolutionary analyses for better explaining adaptations to industry and suggesting genes for genetic engineering have been successfully used in yeast [7,11,21,57,77]. Understanding its evolutionary adaptations may help not only to solidify this bacterium as a potential substitute to the yeast *Saccharomyces cerevisiae* in industry, but also bring novel genes and information that can be used for yeast, enhance its fermenting capabilities, and increase production of our current bio-platforms.

Regarding again xylose transport, most known transporters are inhibited by

glucose or other hexoses, and preferentially carry these sugars instead of xylose during co-fermentation conditions. However, while yeast only have MFS transporters, which are proteins capable of carrying solutes passively through the membrane, bacteria have an additional class called ABC transporters, which are usually organized in operons, that require the use of ATP molecules to pump solutes inward [13,42]. In *E. coli*, the ABC transport system responsible for specifically uptaking xylose is the XylFGH operon, comprised of proteins XylF (external protein responsible for sugar ligation), XylG (ATP-binding) and XylH (translocation of xylose through the membrane), with XylR acting as a regulator. Even though an energy expenditure is needed for consuming xylose, this could confer an advantage when sugar availability is scarce, allowing organisms with these systems to continue thriving in their environment [13,67].

T. sac being a lignocellulosic fermenter with low glucose inhibition could have adaptations in its transporters to efficiently carry xylose, specifically, on the ABC transporters because of their energy consumption load. In this work comparative genomics was used across 20 bacterial species, to identify evolutionary marks related to xylose fermentation. Positive selection was found on the xylF homolog, the sugar binding molecule on ABC transporter, where the largest affinity is needed for xylose transport is shown. Together with adaptations in transport proteins, other T.sac proteins were found to carry important adaptations putatively related to the xylose metabolism. Understanding all these T.sac adaptations may further help alleviate uptake difficulties on current cell factories such as yeast.

2.3 Materials and Methods

2.3.1 Dataset

20 bacterial genomes, including *T.sac*, were downloaded from NCBI for analysis. Genomes were chosen based on proximity to *T.sac* (close and distant species), xylose metabolism capacity and industrial applications (table 1).

Table 1. Information on the used public bacterial genomes obtained on NCBI

Bacteria Name	**Abbreviation**	**Xylose metabolism**	**Assemby ID**	**Reference**
Salmonella enterica	CAD	No	GCA_000195995.1	[53]
Escherichia coli	AAC	Yes	GCA_000005845.2	[6]
Clostridium acetobutylicum	AAK	No	GCA_000008765.1	[52]
Bacillus subtilis	CAB	No	GCA_000009045.1	[39]
Lactobacillus fermentum	BAG	No	GCA_000010145.1	[49]
Lactobacillus acidophilus	AAV	No	GCA_000011985.1	[4]
Clostridium botulinum	ABS	No	GCA_000017045.1	[64]
Natranaerobius thermophilus	ACB	No	GCA_000020005.1	[82]
Ruminiclostridium (Clostridium) celullolyticum	ACL	Yes	GCA_000022065.1	[27]
Thermoanaerobacterium thermosaccharolyticum	AFK	Yes	GCA_000145615.1	[27]
Halanaerobium praevalens	ADO	No	GCA_000165465.1	[31]
Halanaerobium hydrogeniformans	ADQ	Yes	GCA_000166415.1	[9]
Bacillus cellulosilyticus	ADU	No	GCA_000177235.2	[48]
Thermoanaerobacterium xylanolyticum	AEF	Yes	GCA_000189775.3	[41]
Thermoanaerobacterium saccharolyticum	AFK	Yes	GCA_000307585.2	[15]
Thermoanaerobacterium aotearoense	ETO	Yes	GCA_000512105.1	[2]
Bacillus beveridgei	AOM	No	GCA_001721685.1	[5]
Natranaerobius trueperi	OWZ	No	GCA_002216005.1	[24]
Clostridium chauvoei	SLK	No	GCA_900168365.1	[60]
Thermus thermophilus	YP	No	GCF_000091545.1	Masui et al

2.3.2 Orthology assignment

Comparing genes across species requires clustering of their proteins into putative homologous groups. Orthofinder [18] was used to accomplish such task, by first clustering genes according to similarity using DIAMOND [10], and cutoff with the default MCL inflation. Briefly, FASTA files of the translated CDS from each species were used as input for DIAMOND, which is used for finding the most similar sequences via reciprocal-best-hits. Orthofinder, by using the MCL parameter, aggregates the most similar sequences into Orthogroups (gene families, a set of genes from multiple species descended from a single gene).Orthogroups of sugar transporter genes were recovered using BLASTP with known protein sequences from E. coli or T.sac and an e-value threshold of 1e-20, and also. through HMM profiles from Pfam (PF07690.15, PF13347.5 and PF05631.13). Genes and protein names were kept as the NCBI identifier. Family names were kept as outputted from Orthofinder.

2.3.3 Phylogenetic inferences

For each gene family, multiple sequence alignment (MSA) was done using protein sequences through MAFFT [34], anchoring first via local alignment (localpair) and 1000 iterations (maxiterate 1000). Phylogenetic relationships for each transporter family's genes was inferred through Maximum Likelihood, implemented on IQTree [71]. IQTree's automatic model selection tool was used for fitting the best substitution model and branch support was tested by 1000 ultrafast bootstraps. Species' phylogenomics was inferred by concatenating all 157 Single-Copy Orthogroups MSAs with FASConcat [37] and analyzing with IQTree using the same strategy as above and through Bayesian methods implemented via MrBayes [58]. MrBayes was ran using a GTR+ invgamma substitution model with 3 hot and 1 cold chains for 10 million generations discarding 25% of the generations from the cold chain

2.3.4 Gene duplication analysis

Gene birth and death estimation was obtained by using the same Orthofinder gene count output and the ultrametric species tree as inputs for CAFE [16] analysis, which models the evolution of Orthogroups based on the species' phylogeny. CAFE was ran with default parameters, and λ to maximize the log likelihood for all families (lambda -s). Number of expanding (gene birth), contracting (gene loss), average expansion and log of average expansion can be seen on table 2.

Table 2. CAFE results for expanding and retracting gene families. EF means number of Expanding Families and CF represents Contracting Families. For both columns, stat means number of statistically significant families ($p < 0.05$)

Species	EF (stat)	CF (stat)	Avg. Expansion	Log Avg Expansion
Escherichia coli	198 (3)	226 (1)	0.00245	-2.61079
Salmonella enterica	225 (1)	253 (1)	0.001685	-2.77352
Bacillus beveridgei	188 (5)	845 (2)	-0.09081	-1.04186
Bacillus cellulosilyticus	293 (5)	390 (1)	0.0049	-2.30976
Bacillus pseudofirmus	338 (2)	780 (0)	-0.04043	-1.39331
Bacillus subtilis	380 (4)	1157 (1)	-0.0781	-1.10734
Lactobacillus acidophilus	124 (1)	367 (4)	-0.03277	-1.4845
Lactobacillus fermentum	108 (5)	278 (2)	-0.01807	-1.74303
Clostridium chaouvoei	194 (3)	1158 (1)	-0.1415	-0.84924
Clostridium acetobutylicum	308 (7)	554 (0)	-0.00168	-2.77352
Clostridium botulinum	279 (3)	639 (1)	-0.03813	-1.41871
Clostridium cellulolyticum	283 (11)	1329 (0)	-0.12021	-0.92004
Thermoanaerobacterium xylanolyticum	61 (1)	587 (0)	-0.07672	-1.11508
Thermoanaerobacterium saccharolyticum	30 (1)	16 (0)	0.003216	-2.49269
Thermoanaerobacterium aotearoense	20 (0)	25 (0)	-0.00061	-3.21285
Thermoanaerobacterium thermosaccharolyticum	143 (0)	579 (0)	-0.06049	-1.21832
Natranaerobius trueperi	104 (4)	357 (2)	-0.03247	-1.48858
Natranaerobius thermophilus	205 (4)	190 (1)	0.01853	-1.73213
Halanaerobium hydrogeniformentans	179 (7)	345 (1)	-0.01363	-1.86552
Halanaerobium praevalens	94 (3)	386 (1)	-0.04196	-1.37716
Thermus thermophilus	149 (1)	5330 (0)	-0.79387	-0.10025

Sometimes, a simpler approach can give us interesting insights into genome evolutionary dynamics that may be lost to canonical approaches such as CAFE. Thus, besides this model-based analysis, we carried out a naïve gene duplication analysis by analyzing Orthofinder's gene count output using a custom python script, by filtering from the dataset gene families where T.sac had a larger gene count than the mean plus two standard deviations. As retained gene duplications are expected to be scarce, gene families with at least three genes were considered for further detailed examination. Figure 3 shows the results for this analysis.

2.3.5 Natural selection analysis

Nucleotide codon alignments were obtained using MACSE to align the nucleotide sequences by using the protein alignment for each Orthogroup, done through MAFFT (localpair, maxiterate 1000) as template. Codon-based alignments and ML trees for each gene family obtained in OrthoFinder were used in tests for dN/dS models. Selection analysis was implemented using CODEML from the PAML package [80] through ETE3 Toolkit, and MEME from the HYPHY package [51]. Both analyses implement branch-site tests, comparing substitution rates between indicated branches (Foreground) against the other branches (Background). T. sac branch tips were marked for both analysis as Foreground for comparison against the other branches. Models ran for CODEML were bsA and bsA1. Positive selection was inferred for sites with p-value of LRT between alternative model and null model (bsA and bsA1, respectively), if the p-value was smaller than 0.05 the site was retrieved as a positively selected site [81]. Families with positive selection detected by CODEML were subjected to MEME to check if a second method would corroborate the CODEML results. The selected sites were retrieved directly by looking if there were branches under selection at a p-value smaller than 0.05.

2.3.6 Functional annotation

Families with positive selection evidence from MEME and CODEML or with duplications were retrieved for further inquiries and discussion. Functional annotation was retrieved from each protein's NCBI accession, and each family was annotated manually based on the most represented annotation within all proteins. All families with selection were also submitted to Eggnog [30] and PANNZER2 [69] servers as an attempt to better annotate cases in which the original annotation from NCBI was too vague. Briefly, Eggnog uses precomputed Orthogroups with functional description from its database to retrieve the most likely functional information from the inputted sequences. PANNZER2 similarly searches for homolog sequences in the Uniprot database and based on the most similar sequences annotates the most probable function. The most represented annotation within proteins was kept as the family's annotation.

2.3.7 Microarray analysis

Currie et al [15] published microarray data from T. sac fermentation on a variety of conditions, including co-consumption of glucose and xylose, and shocking with hemicellulose (which they called 'washate'). This available data was filtered for proteins of interest obtained from our evolutionary analyses and plotted as a heatmap to help visualize genes with higher expression on these conditions.

2.4 Results

2.4.1 Orthology assignment and phylogenetic inferences

Orthofinder clustering resulted in 14,714 gene families with 2292 families containing at least one T. sac member. Orphan genes (genes not assigned to any family) varied from 1 to 20% for species used (Supplementary figure 1) showing that clustering was satisfactory. A higher orphan rate for some species than others can be explained by the wide phylogenetic range of genomes chosen for analysis.

T. sac gene copies for each family were aligned to check if each copy may have diverged since duplication. All sequences showed nucleotide substitutions in alignment, some being much more fragmented, which might indicate pseudogenization, while others maintained much of their structure showing only single nucleotide substitutions.

Both Bayesian and Maximum Likelihood inferences using 159 single copy gene families reconstructed the same relationships among species (figure 1). One interesting finding is that *T. sac* and *T. aotearoense* had no branching between them and showed no nucleotide differences for the 159 single copy genes, which indicates that they are most likely the same phylogenetic lineage and the same species. A reasonable explanation for why this result differs from what is described in literature is because taxonomic classification of these bacteria was done through 16S [45], which is known not to be always a reliable marker for separating closely related species. A newer revision on GDTB [54] also considers *T. aotearoense* as *T. saccharolyticum*.

2.4.2 Genome-wide evolution and adaptation clues

The approach of estimating gene birth and losses throughout the genome showed that, on average, we have a tendency for gene losses for all genomes with no relationship to the xylose fermentation phenotype (figure 1 and table 2). Despite almost no expansion seen for T. sac, table 2 shows one statistically significant expansion, which is a rapidly evolving family, found to be OG55, annotated as a transposase IS116/IS110/IS902 family protein. When analyzing T. sac family members' positions in the genome, AFK85526.1 and AFK85527.1 were of special interest, as they appear close to each other, with AFK85525.1

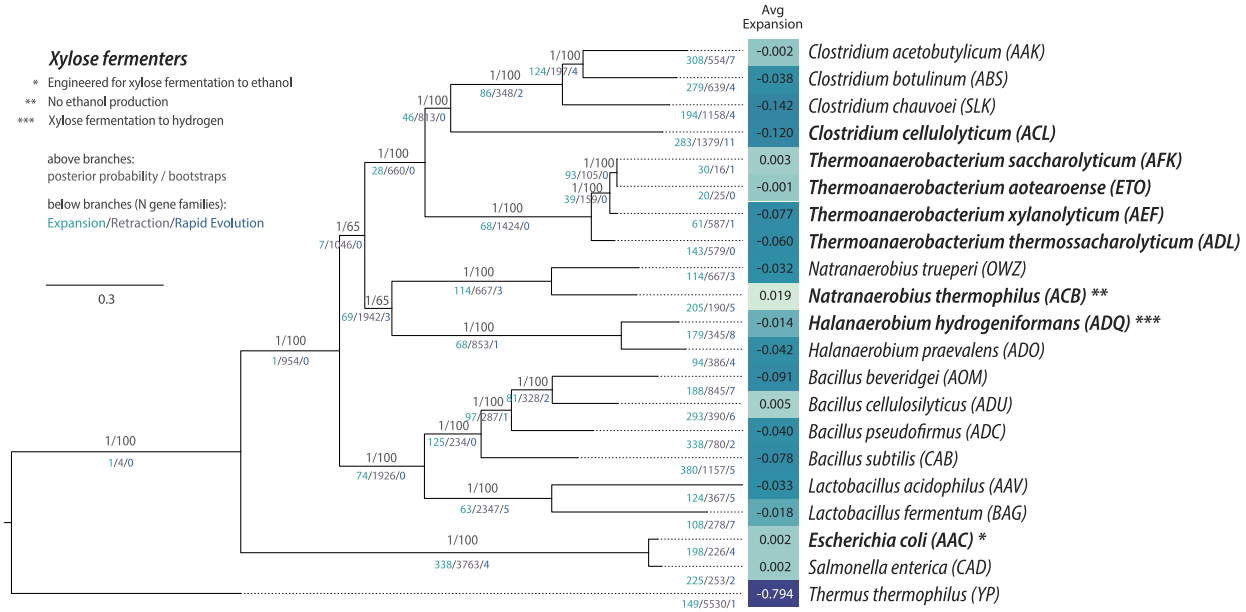


Fig 1. Phylogenetic inference of the 157 single-copy orthogroups for the 20 Firmicutes with *T. thermophilus* as outgroup, through Maximum Likelihood and Bayesian approaches coupled with gene birth and death results. Numbers above branches represent posterior probabilities and bootstraps from MrBayes and IQTree, respectively, and below branches the number of gene families under expansion, retraction or rapid evolution as reported by CAFE. Average expansion was also outputted by CAFE analysis, darker colors represent less expansion (more retraction) while lighter colors represent positive or near zero expansion. Abbreviations after species names represent their NCBI abbreviation which are also used on other figures. Species in bold are able to ferment xylose.

(annotated as a Xylose isomerase domain protein) and AFK85528.1 (annotated as araC) as the closest genes, which are related to pentose metabolism. Besides, this transposase also showed sites under positive selection, as seen both in MEME and CODEML (figure 2).

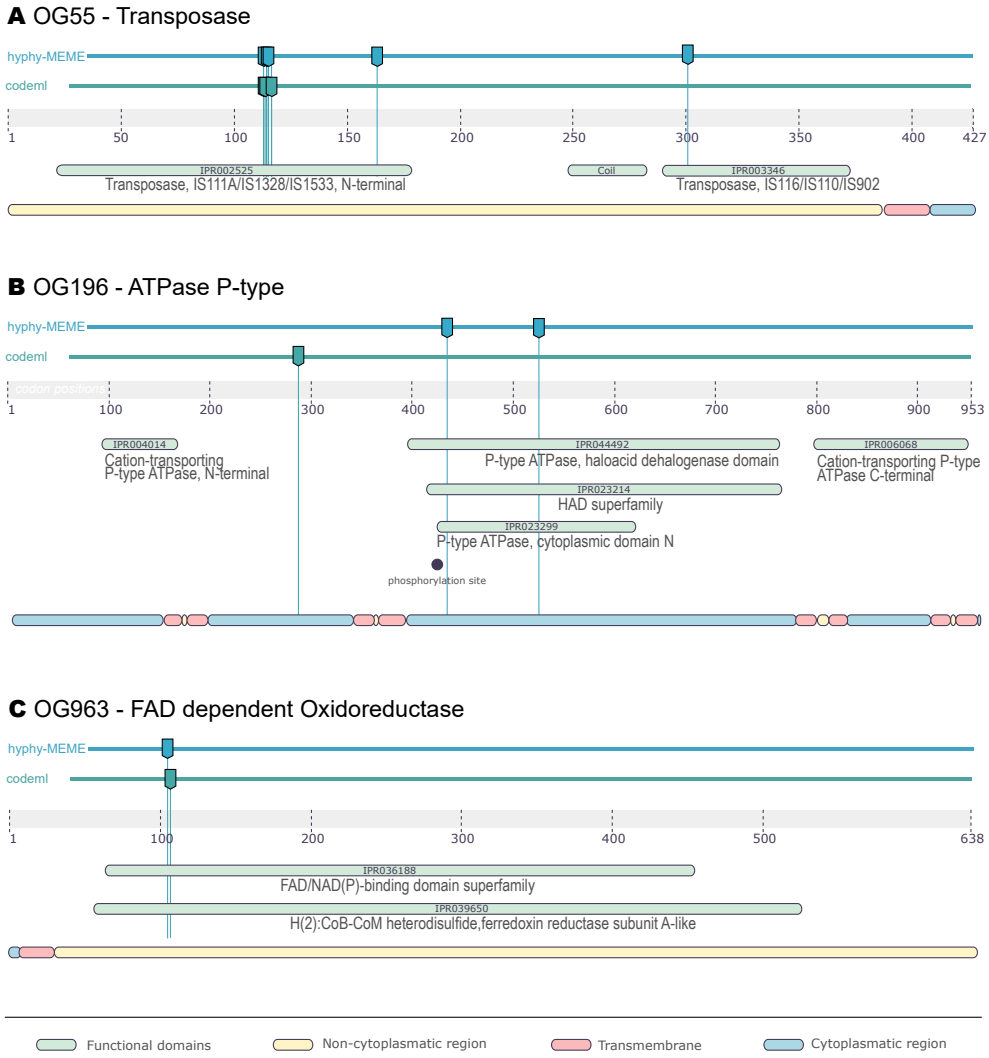


Fig 2. Graphical representation of a *T. sac* protein sequence from families (a) OG55, (b) OG106 and (c) OG963, respectively, with Interpro predicted sites and domains, and sites under positive selection as predicted by HYPHY-MEME and CODEML

To better understand this family's dynamics with these C5 sugar related proteins, we concatenated all xylose isomerase domain families and reconstructed their phylogeny, which revealed monophiletic clades for the proteins of families OG1593, OG2274, and OG2848, with some mixed clades for OG1258, OG1843 and OG4121 (supplementary figure 2). Xylose isomerase families showed no gene expansion in T.sac, high sequence conservation, and no positive selection evidence. Both AFK85525.1 and AFK85528.1 were found to have a baseline expression on public microarray data of *T. sac* on xylose (figure 5). Additionally, AFK85525.1 was found in OG4121, a small family containing only *T. sac*, *T. aotearoense* and *Ruminiclostridium (Clostridium) celullolyticum*, three xylose consuming species.

Regarding positive selection analysis, there were 7 gene families under selection found by CODEML, but only 4 were detected by MEME and thus, we chose to further explore only these four families with coinciding signals: OG55, OG106, OG963 and OG1742

(detailed at a later section). Figure 2 shows OG55, OG106 and OG963 sites with evidence of positive selection, as well as InterproScan [8] predicted domains. The three families showed sites with positive selection within protein domains. Family OG963 was annotated as fumarate reductase/succinate dehydrogenase. Family OG106 comprises HAD superfamily P-type ATPases (PMCA), with a role in translocating calcium, sodium and hydrogen ions across the plasma membrane (Type II subfamily [12]), reducing environmental stress caused by these ions; some downregulation was seen on the microarray xylose conditions for the genes in this family.

2.4.3 Genes possibly related to sugar metabolism with increased gene copies

T. sac xylose and glucose co-fermenting phenotype is remarkable and much desired for industrial applications, thus having a bigger viability and fitness against competition, while resisting abiotic stresses, is essential. One interesting family that was found in the analyses was Peptidoglycan Binding Domain 1 (OG1496), which had more copies when compared with other species in this family through the naïve gene duplication approach (3 copies for *T. sac* against the mean of 0.53 copies for the family). This family is involved with cell wall degradation, such as autolytic lysozymes or cleaving autopeptidases, which may suggest a fluidity and dynamic modulation of T.sac cell wall to resist different stresses [17,22,61].

Another interesting family that followed a pattern of greater copies in T. sac than other species was PH1107 glycosidase (OG1224), related to degradation of glycoside bonds of carbohydrates, such as xylan and other complex sugars, which can help to explain its adaptations to ferment these higher carbohydrates directly, even more so with family member AFK87323.1 being slightly upregulated on all conditions when compared to other members of this Orthogroup.

Acetoin dehydrogenase regulator family (OG1421) was surprising as having 5 copies in T.sac, while the family mean number of copies was 0.62. These genes are responsible for the regulation of acetyl-CoA and acetaldehyde through the reaction

$$acetoin + CoA + NAD+ \rightleftharpoons acetaldehyde + acetyl - CoA + NADH + H$$

and require thiamine diphosphate as a cofactor. The annotation of a LuxR motif is reported in acetoin regulator for other species, such as *Klebsiella pneumoniae* [28,55].

Pyruvate/Ketoisovalerate oxireductase family (OG1052) was interesting, especially *T. sac* protein AFK87181.1 and AFK86082.1 as this group was also detected on our naïve duplication approach and these members were shown to be expressed in xylose and xylan fermenting conditions when looking at the microarray data (glucose-xylose

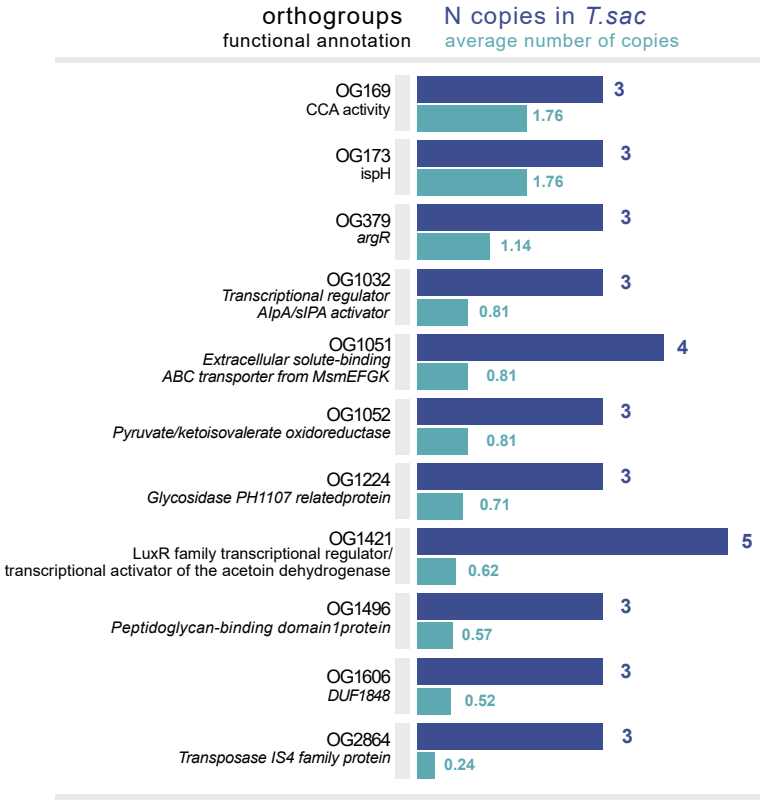


Fig 3. Naïve duplication analysis for gene families in which *T. sac* has more copies than the family mean plus two standard deviations number of copies, and with at least 3 duplications in *T.sac*. Shown in purple are the copies in *T.sac* and in green the mean number of copies for the family.

and xylose_shock). Pyruvate oxireductase is also called pyruvate:ferredoxin oxidoreductase (PFOR), and is involved in the synthesis of acetyl-CoA from pyruvate and oxidized ferredoxin [23,26].

Regarding other sugars, family OG1051, annotated as an extracellular solute-binding protein for the ABC operon MsmEFGK, was interesting. In *Streptococcus mutans* MsmEFGK is responsible for carrying Melibiose, Raffinose, Stachyose, Isomaltose and Isomaltriose [74]. This high copy-number may be explained by a need to transport different solutes, and as an ABC transport requires energy, adjusting these copies to be specific to each sugar may have been what happened during T. sac evolution. Also, although we usually see regulation confined on the operon's region, it has been proposed that different subunits of an ABC transporter can pair with other ABC components to fulfill their function [66,74]. Figure 3 shows information retrieved from our naïve approach.

Proteins from all these families are shown in the microarray (figure 5), with family OG1224 and OG1421 having some of the most expressed genes for most conditions and family OG1051 being mostly repressed or neutral, which is expected as the members are not directly related to xylose.

2.4.4 Genes of xylose transporter families

Six families were identified as MFS transporters and 7 as ABC transporters. The known xylose MFS transporter from *E. coli* xylE was positioned on OG271 according to our BLAST search, however, no *T. sac* genes were present in this family. Analyzing the BLAST results, no *T. sac* or other Thermoanaerobacterium genes were similar to xylE within our e-value cutoff threshold. The most similar gene was AFK86490.1, with 29.49% identity, annotated as a drug resistance transporter (OG25, which was added for analysis, data not shown). This initial finding suggests that *Thermoanaerobacterium* strategies for xylose transport differ from other groups of bacteria, which may be indicative of their efficiency for 2G fermentation.

OG1742 (xylF family) was recovered upon closer inspection of proteins annotated as sugar transporters in the *T. sac* microarray published by Currie et al [15]. AFK86454.1 had a higher expression during their experiments than other similar extracellular binding proteins or MFS transporters. As this family also was not returned through BLAST with known xylF proteins, this may also suggest a different mechanism or adaptation for active transport of xylose.

No evidence of positive selection was found in MFS transporters. Also, for the ABC xylose transporter operon no xylG, xylH or xylR families had any evidence of positive selection. However, as also previously reported for OG55, OG106 and OG963, OG1742 (xylF) showed evidence for positive selection for both CODEML and MEME analyses. Figure 4 shows the family's phylogeny, part of the MSA and sites under positive selection. One interesting finding is that residue 274 (239 ungapped) is two residues distant of site 272 (237 ungapped), which is annotated as one of the sugar binding sites when analyzed through InterproScan.

2.5 Discussion

Gene duplication is an important phenomenon in eukaryotes as a source of novelty and adaptation [70], however in prokaryotes it is rarely seen, as arisen duplications are costly to maintain [1]. Thus, finding multiple copies of a gene for a given family may reflect adaptation in response to a new environmental pressure or functional specialization within the same environment [8]. A large proportion of gene family retractions were found in the Thermoanaerobacterium branch, a much higher retraction than observed in other species as shown in figure 1. This could be related to the stressful environments in which these organisms are found, as it is known that gene deletions can happen in stressful environments as a consequence of reduced usage, reducing energetic costs to the organism [3]. Even though this retraction was seen, Family OG55 of insertion sequences was found to be under rapid evolution.

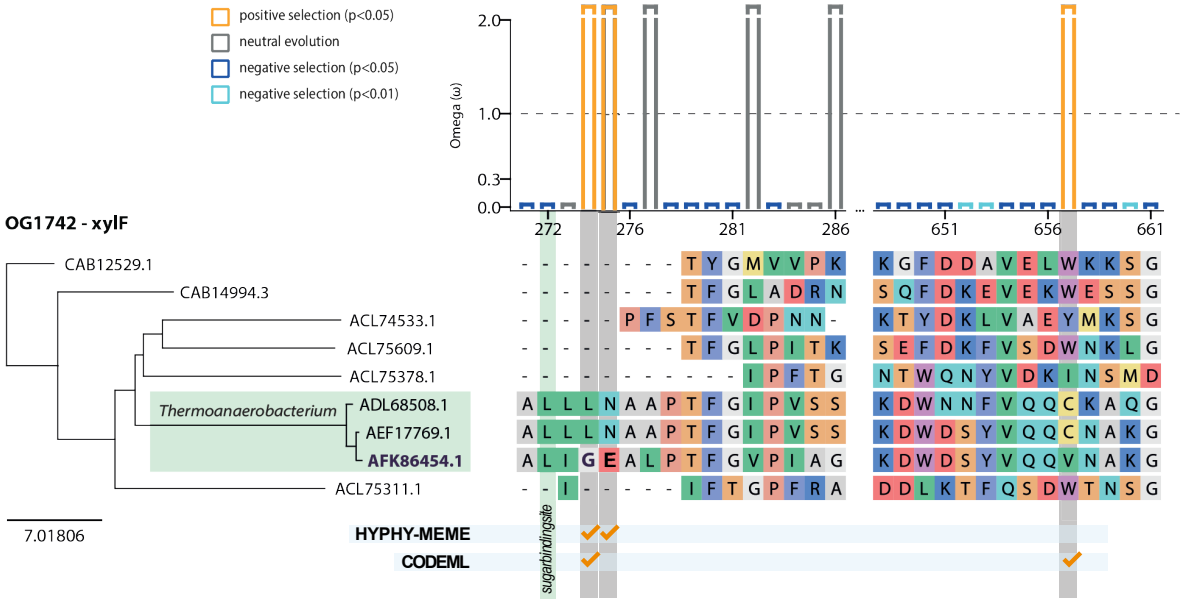


Fig 4. Phylogenetic inference, MSA and sites under positive selection for xylF family OG1742.

Nucleotide substitutions that undergo positive selection through evolution on a given gene, as a consequence of an increase in the organism's fitness, usually leave traces such as an altered rate of substitutions for an amino acid site. Attempting to correlate selection marks with industry desired phenotypes is a newer approach that may help to choose targets for bioengineering of industrial microorganisms [7,21,57,77]. Positive selection clues here revealed three gene families that seem to be important in xylose metabilosm. The OG55 transposase, found to be under positive selection, is positioned side-by-side with a xilose isomerase domain protein and an araC on the T.sac genome. Some studies have also shown that transposable elements may be beneficial by activating metabolism, resistance or acting as a defense mechanism [19,25].

Family OG963 was annotated as Succinate dehydrogenase, which is an oxidoreductase from the complex II electron transport chain. For the 3 genes in this family, an InterproScan search revealed an heterodisulfide reductase domain, which is associated with methanogenic reduction of ferredoxin [33], and in thermophilic bacteria, hydrogen is formed from ferredoxin, enabling extra ATP production [62]. One site in this region was found to be under positive selection, and AFK87395.1 was found to be more expressed in some xylose conditions, as seen in the microarray data. In T. sac, it was also previously described that Ferredoxin:NAD+ Oxidorreductase is essential for Ethanol formation [68].

Regarding families related to sugar metabolism, families OG1224, OG1421 and OG1052 showed a greater number of genes on our naïve approach. Family OG1224 of PH1107 glycosidase is related to degradation of glycoside bonds of carbohydrates, such as xylan and other complex sugars, which can help to explain its adaptations to ferment these carbohydrates directly, even more so with AFK87323.1 being upregulated on

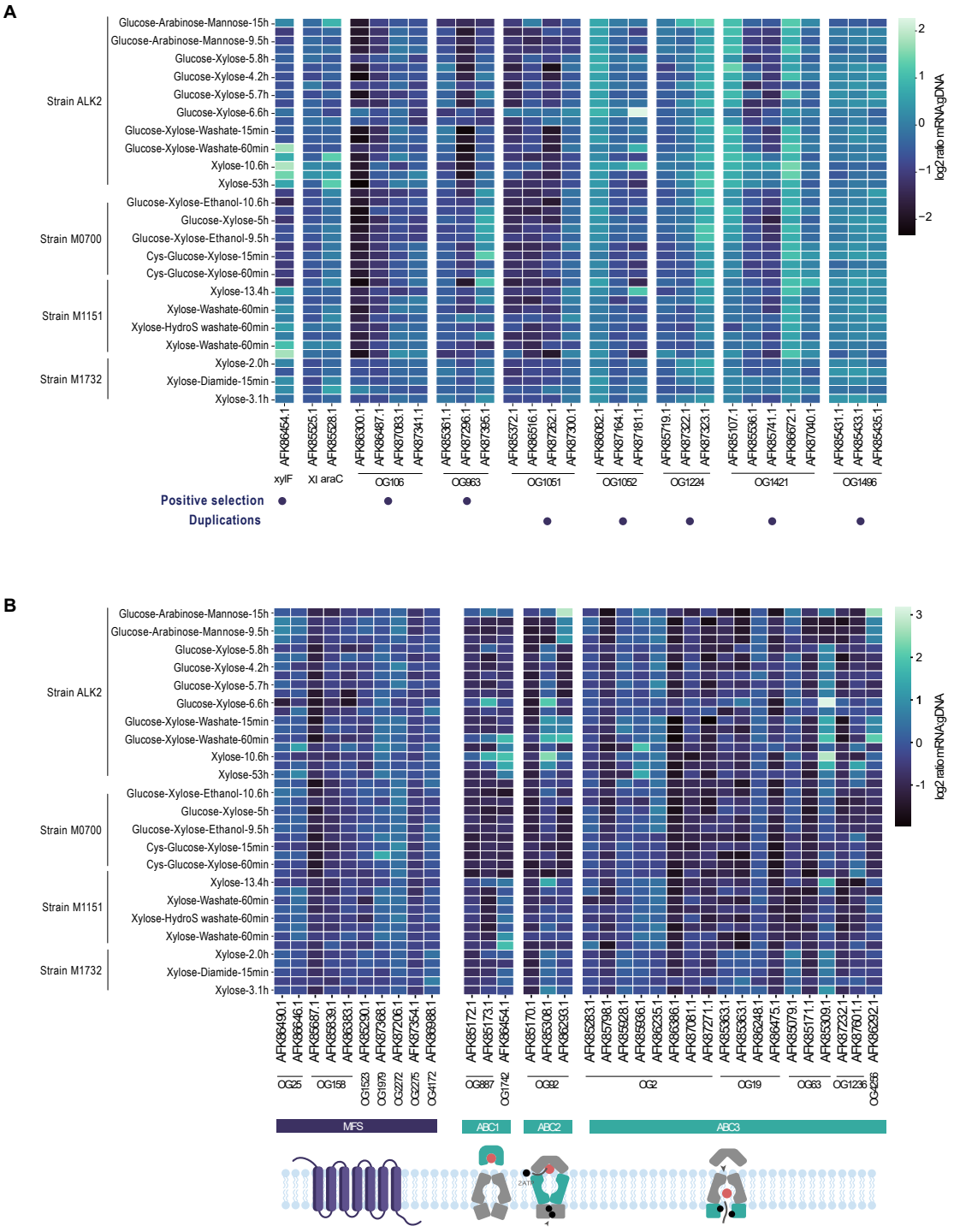


Fig 5. Heatmap adapted from Currie et al microarray data showing (a) xylose related genes with putative adaptations and (b) sugar transporters. Each row of the heatmap is represented as condition – time point.

xylose conditions. Family OG1421 of acetoin dehydrogenase was particularly interesting as AFK86672.1 is upregulated in many xylose conditions as seen in the microarray data; as acetoin can be a fermentation inhibitor, this reaction could be extremely important for cofactor regeneration and disinhibition, especially in stressful conditions [78], furthermore, it has been reported that acetoin can also be used as a carbon source by Bacillus subtilis [29,76]. Regarding OG1052, in T. sac, it has been shown that PFOR is present as at least six clusters, with AFK87181.1 as a member of pforF and AFK86082.1 as a member of pforD. These genes are important for pyruvate dissimilation, with pforA (Tsac_0046/ AFK85084.1; OG320) as the most important member for ethanol production [85]. As pforA is not a member of OG1052, multiple copies of PFOR with slight differences may help alleviate redox needs for the cells and quicken reactions, as hinted by both the microarray data and STRINGdb network [65], which shows that both proteins are co-expressed with 4Fe-4S ferredoxin and other genes of this metabolical pathway.

As mentioned before, one of the bottlenecks in industrial 2G fermentation is inhibition of pentose transport by C6 sugars, thus, finding better xylose transporters is desirable to mitigate this occurrence.

Studies with MFS sugar transporters show that most are found to be under purifying selection [73,79], and better fermenting species are known to have more gene duplications, suggesting that sugar transport evolved in a multi-genic strategy to increase transport speed while maintaining their sequence relatively unchanged [44]. Because of the ATP dependency by ABC transporters, we hypothesized that adaptations should have occurred to compensate the risk of dispending energy to capture sugar molecules. Also, selective pressure signals should appear most likely on the sugar binding protein (xylF) as this energy expenditure would signify specificity to xylose. Thus, we searched for evidence of positive selection on families related to sugar transport, both MFS and ABC. No MFS transporters or xylFGH operon members had evidence of positive selection, except xylF (OG1742), which was found to be positively selected. This is interesting because xylF being the xylFGH operon's extracellular protein, responsible for capturing xylose molecules, and being and ABC transporter, strongly indicates specificity to xylose to compensate the energy expenditure in transporting solutes. Moreover, most adaptations found in sugar transporters for yeast after rounds of directed evolution or mutagenesis are also in amino acid residues close to the predicted binding sites [20,56]. Evidence of positive selection, coupled with higher expression during xylose fermentation as seen on the published microarray suggests a higher adaptation for xylose metabolism in this organism by increasing its efficiency on capturing xylose and may be a key reason on why it is capable to co-ferment glucose and xylose. Not many sugar transporters specific for xylose are known and having an independent route for assimilation may alleviate competition of sugar molecules to the active site of promiscuous MFS transporters.

Regarding the possibility of using this transporter for present microbial cell factories, in eukaryotes there are almost no ABC transporters functioning on importing molecules, only exporters [75]. However, for industrial yeast, in addition to extracellular pumps related to drug removal and detoxification, there are some ABC importers, such as AUS1 and PDR11 which are related to external sterol assimilation for ergosterol production [35,38]. Moreover, horizontal gene transfer has been reported from bacteria to fungi [36,59], including some genes associated with xylose metabolism, such as Xylose Isomerase [50]. This indicates that adapting current *T. sac* xylFGH operon for use in yeast may be a possibility to help co-consumption of xylose and glucose during 2G ethanol fermentation, such as by mutating key residues in MFS transporters using *T. sac* xylF as a model.

Thermoanaerobacter species, such as *Thermoanaerobacterium saccharolyticum*, have several traits that can benefit industrial biotechnological fermentations, especially 2G ethanol production, for which *T. sac* has been engineered to metabolize at a great yield. In addition to genomic and transcriptomic resources already published, a deep evolutionary analysis and understanding of *T. sac* by comparing its genome against other bacteria groups is presented. Genomic adaptations to environmental stress are shown, such as heat stress and iron reduction, as well as xylose metabolism, seen at the specific xylFGH xylose operon, which has been also found in expression data, showing that evolutionary exploratory analysis can be useful for biotechnological prospecting. These data can serve as basis on searching for targets for industrial adaptation of *T. sac*, as well as reveal novel genes for yeast engineering.

2.6 References

[1].Adler, M., Anjum, M., Berg, O.G., Andersson, D.I., Sandegren, L. (2014) High fitness costs and instability of gene duplications reduce rates of evolution of new genes by duplication-divergence mechanisms. Molecular Biology and Evolution 31(6), 1526–35, Doi: 10.1093/molbev/msu111.

[2].Ai, H., Zhang, J., Yang, M., Yu, P., Li, S., Zhu, M., Dong, H., Wang, S., Wang, J. (2014) Draft Genome Sequence of an Anaerobic, Thermophilic Bacterium, Thermoanaerobacterium aotearoense SCUT27, Isolated from a Hot Spring in China. Genome Announcements 2(1), Doi: 10.1128/genomea.00041-14.

[3].Albalat, R., Cañestro, C. (2016) Evolution by gene loss. Nat Rev Genet 17(7), 379–91, Doi: 10.1038/nrg.2016.39.

[4].Altermann, E., Russell, W.M., Azcarate-Peril, M.A., Barrangou, R., Buck, B.L., McAuliffe, O., Souther, N., Dobson, A., Duong, T., Callanan, M., Lick, S., Hamrick, A., Cano, R., Klaenhammer, T.R. (2005) Complete genome sequence of the probiotic lactic acid bacterium Lactobacillus acidophilus NCFM. Proceedings of the National

Academy of Sciences of the United States of America 102(11), 3906–12, Doi: 10.1073/pnas.0409188102.Abstract/FREE Full Text

[5].Baesman, S.M., Stolz, J.F., Kulp, T.R., Oremland, R.S. (2009) Enrichment and isolation of Bacillus beveridgei sp. nov., a facultative anaerobic haloalkaliphile from Mono Lake, California, that respires oxyanions of tellurium, selenium, and arsenic. Extremophiles 13(4), 695–705, Doi: 10.1007/s00792-009-0257-z.

[6].Blattner, F.R., Plunkett, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B., Shao, Y. (1997) The complete genome sequence of Escherichia coli K-12. Science 277(5331), 1453–62, Doi: 10.1126/science.277.5331.1453.Abstract/FREE Full Text

[7].Borelli, G., Fiamenghi, M.B., Dos Santos, L.V., Carazzolle, M.F., Pereira, G.A.G., José, J., Zhang, G. (2019) Positive Selection Evidence in Xylose-Related Genes Suggests Methylglyoxal Reductase as a Target for the Improvement of Yeasts' Fermentation in Industry. Genome Biology and Evolution 11(7), 1923–38, Doi: 10.1093/gbe/evz036.

[8].Bratlie, M.S., Johansen, J., Sherman, B.T., Huang, D.W., Lempicki, R.A., Drabløs, F. (2010) Gene duplications in prokaryotes can be associated with environmental adaptation. BMC Genomics 11(1), Doi: 10.1186/1471-2164-11-588. [9].Brown, S.D., Begemann, M.B., Mormile, M.R., Wall, J.D., Han, C.S., Goodwin, L.A., Pitluck, S., Land, M.L., Hauser, L.J., Elias, D.A. (2011) Complete genome sequence of the haloalkaliphilic, hydrogen-producing bacterium Halanaerobium hydrogeniformans. Journal of Bacteriology 193(14), 3682–3, Doi: 10.1128/JB.05209-11.Abstract/FREE Full Text

[10].Buchfink, B., Xie, C., Huson, D.H. (2014) Fast and sensitive protein alignment using DIAMOND. Nature Methods 12(1), 59–60, Doi: 10.1038/nmeth.3176.

[11].Bueno, J.G.R., Borelli, G., Corrêa, T.L.R., Fiamenghi, M.B., José, J., de Carvalho, M., de Oliveira, L.C., Pereira, G.A.G., dos Santos, L.V. (2020) Novel xylose transporter Cs4130 expands the sugar uptake repertoire in recombinant Saccharomyces cerevisiae strains at high xylose concentrations. Biotechnology for Biofuels 13(1), 145, Doi: 10.1186/s13068-020-01782-0.

[12].Burroughs, A.M., Allen, K.N., Dunaway-Mariano, D., Aravind, L. (2006) Evolutionary Genomics of the HAD Superfamily: Understanding the Structural Adaptations and Catalytic Diversity in a Superfamily of Phosphoesterases and Allied Enzymes. Journal of Molecular Biology 361(5), 1003–34, Doi: 10.1016/j.jmb.2006.06.049.

[13].Cui, J., Davidson, A.L. (2011) ABC solute importers in bacteria. Essays in Biochemistry 50(1), 85–99, Doi: 10.1042/BSE0500085.Abstract/FREE Full Text

[14].Cui, J., Maloney, M.I., Olson, D.G., Lynd, L.R. (2020) Conversion of

phosphoenolpyruvate to pyruvate in Thermoanaerobacterium saccharolyticum. Metabolic Engineering Communications 10, e00122, Doi: 10.1016/j.mec.2020.e00122.

[15].Currie, D.H., Raman, B., Gowen, C.M., Tschaplinski, T.J., Land, M.L., Brown, S.D., Covalla, S.F., Klingeman, D.M., Yang, Z.K., Engle, N.L., Johnson, C.M., Rodriguez, M., Joe Shaw, A., Kenealy, W.R., Lynd, L.R., Fong, S.S., Mielenz, J.R., Davison, B.H., Hogsett, D.A., Herring, C.D. (2015) Genome-scale resources for Thermoanaerobacterium saccharolyticum. BMC Systems Biology 9(1), 1–15, Doi: 10.1186/s12918-015-0159-x.

[16].De Bie, T., Cristianini, N., Demuth, J.P., Hahn, M.W. (2006) CAFE: A computational tool for the study of gene family evolution. Bioinformatics 22(10), 1269–71, Doi: 10.1093/bioinformatics/btl097.

[17].Dideberg, O., Charlier, P., Dive, G., Joris, B., Frère, J.M., Ghuysen, J.M. (1982) Structure of a Zn 2+ - containing D -alanyl-D -alanine-cleaving carboxypeptidase at 2.5 Å resolution. Nature 1982 299:5882 299(5882), 469–70, Doi: 10.1038/299469a0.

[18].Emms, D.M., Kelly, S. (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biology 16(1), 157, Doi: 10.1186/s13059-015-0721-2.

[19].Fan, C., Wu, Y.H., Decker, C.M., Rohani, R., Gesell Salazar, M., Ye, H., Cui, Z., Schmidt, F., Huang, W.E. (2019) Defensive Function of Transposable Elements in Bacteria. ACS Synthetic Biology 8(9), 2141–51, Doi: 10.1021/acssynbio.9b00218.

[20].Farwick, A., Bruder, S., Schadeweg, V., Oreb, M., Boles, E. (2014) Engineering of yeast hexose transporters to transport D-xylose without inhibition by D-glucose. Proceedings of the National Academy of Sciences of the United States of America 111(14), 5159–64, Doi: 10.1073/pnas.1323464111.Abstract/FREE Full Text

[21].Fiamenghi, M.B., Bueno, J.G.R., Camargo, A.P., Borelli, G., Carazzolle, M.F., Pereira, G.A.G., dos Santos, L.V., José, J. (2022) Machine learning and comparative genomics approaches for the discovery of xylose transporters in yeast. Biotechnology for Biofuels and Bioproducts 15(1), 57, Doi: 10.1186/s13068-022-02153-7.

[22].Foster, S.J. (1991) Cloning, expression, sequence analysis and biochemical characterization of an autolytic amidase of Bacillus subtilis 168 trpC2. Microbiology 137(8), 1987–98, Doi: 10.1099/00221287-137-8-1987.

[23].Furdui, C., Ragsdale, S.W. (2000) The role of pyruvate ferredoxin oxidoreductase in pyruvate synthesis during autotrophic growth by the Wood-Ljungdahl pathway. Journal of Biological Chemistry 275(37), 28494–9, Doi: 10.1074/jbc.M003291200.Abstract/FREE Full Text

[24].Guo, X., Liao, Z., Holtzapple, M., Hu, Q., Zhao, B. (2017) Draft genome sequence of natranaerobius trueperi DSM 18760T, an anaerobic, halophilic, alkaliphilic,

thermotolerant bacterium isolated from a soda lake. Genome Announcements 5(36), Doi: 10.1128/genomeA.00785-17.Abstract/FREE Full Text

[25].Hall, B.G. (1998) Activation of the bgl operon by adaptive mutation. Molecular Biology and Evolution 15(1), 1–5, Doi: 10.1093/oxfordjournals.molbev.a025842.

[26].Heider, J., Mai, X., Adams, M.W.W. (1996) Characterization of 2-ketoisovalerate ferredoxin oxidoreductase, a new and reversible coenzyme A-dependent enzyme involved in peptide fermentation by hyperthermophilic archaea. Journal of Bacteriology 178(3), 780–7, Doi: 10.1128/jb.178.3.780-787.1996.Abstract/FREE Full Text

[27].Hemme, C.L., Mouttaki, H., Lee, Y.J., Zhang, G., Goodwin, L., Lucas, S., Copeland, A., Lapidus, A., Del Rio, T.G., Tice, H., Saunders, E., Brettin, T., Detter, J.C., Han, C.S., Pitluck, S., Land, M.L., Hauser, L.J., Kyrpides, N., Mikhailova, N., He, Z., Wu, L., Van Nostrand, J.D., Henrissat, B., He, Q., Lawson, P.A., Tanner, R.S., Lynd, L.R., Wiegel, J., Fields, M.W., Arkin, A.P., Schadt, C.W., Stevenson, B.S., McInerney, M.J., Yang, Y., Dong, H., Xing, D., Ren, N., Wang, A., Huhnke, R.L., Mielenz, J.R., Ding, S.Y., Himmel, M.E., Taghavi, S., Van Der Lelie, D., Rubin, E.M., Zhou, J. (2010) Sequencing of multiple clostridial genomes related to biomass conversion and biofuel production. Journal of Bacteriology 192(24), 6494–6, Doi: 10.1128/JB.01064-10.Abstract/FREE Full Text

[28].Hsu, J.-L., Peng, H.-L., Chang, H.-Y. (2008) The ATP-binding motif in AcoK is required for regulation of acetoin catabolism in Klebsiella pneumoniae CG43. Biochemical and Biophysical Research Communications 376(1), 121–7, Doi: 10.1016/j.bbrc.2008.08.103.

[29].Huang, M., Oppermann-Sanio, F.B., Steinbüchel, A. (1999) Biochemical and Molecular Characterization of the Bacillus subtilis Acetoin Catabolic Pathway. Journal of Bacteriology 181(12), 3837–41, Doi: 10.1128/JB.181.12.3837-3841.1999.Abstract/FREE Full Text

[30].Huerta-Cepas, J., Forslund, K., Coelho, L.P., Szklarczyk, D., Jensen, L.J., von Mering, C., Bork, P. (2017) Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. Molecular Biology and Evolution 34(8), 2115–22, Doi: 10.1093/molbev/msx148.

[31].Ivanova, N., Sikorski, J., Chertkov, O., Nolan, M., Lucas, S., Hammon, N., Deshpande, S., Cheng, J.F., Tapia, R., Han, C., Goodwin, L., Pitluck, S., Huntemann, M., Liolios, K., Pagani, I., Mavromatis, K., Ovchinikova, G., Pati, A., Chen, A., Palaniappan, K., Land, M., Hauser, L., Brambilla, E.M., Kannan, K.P., Rohde, M., Tindall, B.J., Göker, M., Detter, J.C., Woyke, T., Bristow, J., Eisen, J.A., Markowitz, V., Hugenholtz, P., Kyrpides, N.C., Klenk, H.P., Lapidus, A. (2011) Complete genome sequence of the extremely Halophilic Halanaerobium praevalens type strain (GSL T). Standards in Genomic Sciences 4(3), 312–21, Doi: 10.4056/sigs.1824509.

[32].Joe Shaw, A., Jenney, F.E., Adams, M.W.W.W., Lynd, L.R. (2008) End-

product pathways in the xylose fermenting bacterium, Thermoanaerobacterium saccharolyticum. Enzyme and Microbial Technology 42(6), 453–8, Doi: 10.1016/j.enzmictec.2008.01.005.

[33].Kaster, A.K., Moll, J., Parey, K., Thauer, R.K. (2011) Coupling of ferredoxin and heterodisulfide reduction via electron bifurcation in hydrogenotrophic methanogenic archaea. Proceedings of the National Academy of Sciences of the United States of America 108(7), 2981–6, Doi: 10.1073/pnas.1016761108.Abstract/FREE Full Text

[34].Katoh, K., Standley, D.M. (2013) MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution 30(4), 772–80, Doi: 10.1093/molbev/mst010.

[35].Kohut, P., Wüstner, D., Hronska, L., Kuchler, K., Hapala, I., Valachovic, M. (2011) The role of ABC proteins Aus1p and Pdr11p in the uptake of external sterols in yeast: Dehydroergosterol fluorescence study. Biochemical and Biophysical Research Communications 404(1), 233–8, Doi: 10.1016/j.bbrc.2010.11.099.

[36].Kominek, J., Doering, D.T., Opulente, D.A., Shen, X.-X., Zhou, X., DeVirgilio, J., Hulfachor, A.B., Groenewald, M., Mcgee, M.A., Karlen, S.D., Kurtzman, C.P., Rokas, A., Hittinger, C.T. (2019) Eukaryotic Acquisition of a Bacterial Operon. Cell 176(6), 1356-1366.e10, Doi: 10.1016/j.cell.2019.01.034.

[37].Kück, P., Longo, G.C. (2014) FASconCAT-G: Extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. Frontiers in Zoology 11(1), 81, Doi: 10.1186/s12983-014-0081-x.

[38].Kumari, S., Kumar, M., Gaur, N.A., Prasad, R. (2021) Multiple roles of ABC transporters in yeast. Fungal Genetics and Biology 150(January), 103550, Doi: 10.1016/j.fgb.2021.103550.

[39].Kunst, F., Ogasawara, N., Moszer, I., Albertini, A.M., Alloni, G., Azevedo, V., Bertero, M.G., Bessières, P., Bolotin, A., Borchert, S., Borriss, R., Boursier, L., Brans, A., Braun, M., Brignell, S.C., Bron, S., Brouillet, S., Bruschi, C.V., Caldwell, B., Capuano, V., Carter, N.M., Choi, S.K., Codani, J.J., Connerton, I.F., Cummings, N.J., Daniel, R.A., Denizot, F., Devine, K.M., Düsterhöft, A., Ehrlich, S.D., Emmerson, P.T., Entian, K.D., Errington, J., Fabret, C., Ferrari, E., Foulger, D., Fritz, C., Fujita, M., Fujita, Y., Fuma, S., Galizzi, A., Galleron, N., Ghim, S.Y., Glaser, P., Goffeau, A., Golightly, E.J., Grandi, G., Guiseppi, G., Guy, B.J., Haga, K., Haiech, J., Harwood, C.R., Hénaut, A., Hilbert, H., Holsappel, S., Hosono, S., Hullo, M.F., Itaya, M., Jones, L., Joris, B., Karamata, D., Kasahara, Y., Klaerr-Blanchard, M., Klein, C., Kobayashi, Y., Koetter, P., Koningstein, G., Krogh, S., Kumano, M., Kurita, K., Lapidus, A., Lardinois, S., Lauber, J., Lazarevic, V., Lee, S.M., Levine, A., Liu, H., Masuda, S., Mauël, C., Médigue, C., Medina, N., Mellado, R.P., Mizuno, M., Moestl, D., Nakai, S., Noback, M., Noone, D., O'Reilly, M., Ogawa, K., Ogiwara, A., Oudega, B., Park, S.H., Parro, V., Pohl, T.M., Portetelle, D., Porwollik, S.,

Prescott, A.M., Presecan, E., Pujic, P., Purnelle, B., Rapoport, G., Rey, M., Reynolds, S., Rieger, M., Rivolta, C., Rocha, E., Roche, B., Rose, M., Sadaie, Y., Sato, T., Scanlan, E., Schleich, S., Schroeter, R., Scoffone, F., Sekiguchi, J., Sekowska, A., Seror, S.J., Serror, P., Shin, B.S., Soldo, B., Sorokin, A., Tacconi, E., Takagi, T., Takahashi, H., Takemaru, K., Takeuchi, M., Tamakoshi, A., Tanaka, T., Terpstra, P., Tognoni, A., Tosato, V., Uchiyama, S., Vandenbol, M., Vannier, F., Vassarotti, A., Viari, A., Wambutt, R., Wedler, E., Wedler, H., Weitzenegger, T., Winters, P., Wipat, A., Yamamoto, H., Yamane, K., Yasumoto, K., Yata, K., Yoshida, K., Yoshikawa, H.F., Zumstein, E., Yoshikawa, H., Danchin, A. (1997) The complete genome sequence of the gram-positive bacterium Bacillus subtilis. Nature 390(6657), 249–56, Doi: 10.1038/36786.

[40].Kwak, S., Jin, Y.-S. (2017) Production of fuels and chemicals from xylose by engineered Saccharomyces cerevisiae: a review and perspective. Microbial Cell Factories 16(1), 82, Doi: 10.1186/s12934-017-0694-9.

[41].Lee, Y.E., Jain, M.K., Lee, C., Lowe, S.E., Zeikus, J.G. (1993) Taxonomic distinction of saccharolytic thermophilic anaerobes. International Journal of Systematic Bacteriology 43(1), 41–51, Doi: 10.1099/00207713-43-1-41.

[42].Lewinson, O., Livnat-Levanon, N. (2017) Mechanism of Action of ABC Importers: Conservation, Divergence, and Physiological Adaptations. Journal of Molecular Biology 429(5), 606–19, Doi: 10.1016/j.jmb.2017.01.010.

[43].Lin, L., Song, H., Tu, Q., Qin, Y., Zhou, A., Liu, W., He, Z., Zhou, J., Xu, J. (2011) The Thermoanaerobacter glycobiome reveals mechanisms of pentose and hexose co-utilization in bacteria. PLoS Genetics 7(10), e1002318, Doi: 10.1371/journal.pgen.1002318.

[44].Lin, Z., Li, W.H. (2011) Expansion of Hexose Transporter Genes Was Associated with the Evolution of Aerobic Fermentation in Yeasts. Molecular Biology and Evolution 28(1), 131–42, Doi: 10.1093/MOLBEV/MSQ184.

[45].Liu, S.-Y., Rainey, F.A., Morgan, H.W., Mayer, F., Wiegel, J. (1996) Thermoanaerobacterium aotearoense sp. nov., a Slightly Acidophilic, Anaerobic Thermophile Isolated from Various Hot Springs in New Zealand, and Emendation of the Genus Thermoanaerobacterium. International Journal of Systematic Bacteriology 46(2), 388–96, Doi: 10.1099/00207713-46-2-388.

[46].Mai, V., Lorenz, W.W., Wiegel, J. (2006) Transformation of Thermoanaerobacterium sp. strain JW/SL-YS485 with plasmid pIKM1 conferring kanamycin resistance. FEMS Microbiology Letters 148(2), 163–7, Doi: 10.1111/j.1574-6968.1997.tb10283.x.

[47].Mai, V., Wiegel, J. (2000) Advances in development of a genetic system for Thermoanaerobacterium spp.: Expression of genes encoding hydrolytic enzymes, development of a second shuttle vector, and integration of genes into the chromosome. Applied and Environmental Microbiology 66(11), 4817–21, Doi: 10.1128/AEM.66.11.4817-

4821.2000.Abstract/FREE Full Text

[48].Mead, D., Drinkwater, C., Brumm, P.J. (2013) Genomic and Enzymatic Results Show Bacillus cellulosilyticus Uses a Novel Set of LPXTA Carbohydrases to Hydrolyze Polysaccharides. PLoS ONE 8(4), e61131, Doi: 10.1371/journal.pone.0061131.

[49].Morita, H., Hidehiro, T.O.H., Fukuda, S., Horikawa, H., Oshima, K., Suzuki, T., Murakami, M., Hisamatsu, S., Kato, Y., Takizawa, T., Fukuoka, H., Yoshimura, T., Itoh, K., O'Sullivan, D.J., Mckay, L.L., Ohno, H., Kikuchi, J., Masaoka, T., Hattori, M. (2008) Comparative genome analysis of Lactobacillus renteri and Lactobacillus fermentum reveal a genomic Island for reuterin and cobalamin production. DNA Research 15(3), 151–61, Doi: 10.1093/dnares/dsn009.

[50].Murphy, C.L., Youssef, N.H., Hanafy, R.A., Couger, M.B., Stajich, J.E., Wang, Y., Baker, K., Dagar, S.S., Griffith, G.W., Farag, I.F., Callaghan, T.M., Elshahed, M.S. (2019) Horizontal Gene Transfer as an Indispensable Driver for Evolution of Neocallimastigomycota into a Distinct Gut-Dwelling Fungal Lineage. Applied and Environmental Microbiology 85(15), e00988–19, Doi: 10.1128/AEM.00988-19.Abstract/FREE Full Text

[51].Murrell, B., Wertheim, J.O., Moola, S., Weighill, T., Scheffler, K. (2012) Detecting Individual Sites Subject to Episodic Diversifying Selection. PLoS Genet 8(7), 1002764, Doi: 10.1371/journal.pgen.1002764.

[52].Nölling, J., Breton, G., Omelchenko, M.V., Makarova, K.S., Zeng, Q., Gibson, G., Hong Mei Lee, Dubois, J., Qiu, D., Hitti, J., Aldredge, T., Ayers, M., Bashirzadeh, R., Bochner, H., Boivin, M., Bross, S., Bush, D., Butler, C., Caron, A., Caruso, A., Cook, R., Daggett, P., Deloughery, C., Egan, J., Ellston, D., Engelstein, M., Ezedi, J., Gilbert, K., Goyal, A., Guerin, J., Ho, T., Holtham, K., Joseph, P., Keagle, P., Kozlovsky, J., LaPlante, M., LeBlanc, G., Lumm, W., Majeski, A., McDougall, S., Mank, P., Mao, J.I., Nocco, D., Patwell, D., Phillips, J., Pothier, B., Prabhakar, S., Richterich, P., Rice, P., Rosetti, D., Rossetti, M., Rubenfield, M., Sachdeva, M., Snell, P., Spadafora, R., Spitzer, L., Shimer, G., Thomann, H.U., Vicaire, R., Wall, K., Wang, Y., Weinstock, K., Lai Peng Wong, Wonsey, A., Xu, Q., Zhang, L., Wolf, Y.I., Tatusov, R.L., Sabathe, F., Doucette-Stamm, L., Soucaille, P., Daly, M.J., Bennett, G.N., Koonin, E.V., Smith, D.R. (2001) Genome sequence and comparative analysis of the solvent-producing bacterium Clostridium acetobutylicum. Journal of Bacteriology 183(16), 4823–38, Doi: 10.1128/JB.183.16.4823-4838.2001.Abstract/FREE Full Text

[53].Parkhill, J., Dougan, G., James, K.D., Thomson, N.R., Pickard, D., Wain, J., Churcher, C., Mungall, K.L., Bentley, S.D., Holden, M.T.G., Sebaihia, M., Baker, S., Basham, D., Brooks, K., Chillingworth, T., Connerton, P., Cronin, A., Davis, P., Davies, R.M., Dowd, L., White, N., Farrar, J., Feltwell, T., Hamlin, N., Haque, A., Hien, T.T., Holroyd, S., Jagels, K., Krogh, A., Larsen, T.S., Leather, S., Moule, S., Ó'Gaora, P., Parry,

C., Quail, M., Rutherford, K., Simmonds, M., Skelton, J., Stevens, K., Whitehead, S., Barrell, B.G. (2001) Complete genome sequence of a multiple drug resistant Salmonella enterica serovar Typhi CT18. Nature 413(6858), 848–52, Doi: 10.1038/35101607.

[54].Parks, D.H., Chuvochina, M., Chaumeil, P.A., Rinke, C., Mussig, A.J., Hugenholtz, P. (2020) A complete domain-to-species taxonomy for Bacteria and Archaea. Nature Biotechnology 38(9), 1079–86, Doi: 10.1038/s41587-020-0501-8.

[55].Peng, H.L., Yang, Y.H., Deng, W.L., Chang, H.Y. (1997) Identification and characterization of acoK, a regulatory gene of the Klebsiella pneumoniae acoABCD operon. Journal of Bacteriology 179(5), 1497–504, Doi: 10.1128/jb.179.5.1497-1504.1997.Abstract/FREE Full Text

[56].Reider Apel, A., Ouellet, M., Szmidt-Middleton, H., Keasling, J.D., Mukhopadhyay, A. (2016) Evolved hexose transporter enhances xylose uptake and glucose/xylose co-utilization in Saccharomyces cerevisiae. Scientific Reports 6(1), 19512, Doi: 10.1038/srep19512.

[57].Riley, R., Haridas, S., Wolfe, K.H., Lopes, M.R., Hittinger, C.T., Göker, M., Salamov, A.A., Wisecaver, J.H., Long, T.M., Calvey, C.H., Aerts, A.L., Barry, K.W., Choi, C., Clum, A., Coughlan, A.Y., Deshpande, S., Douglass, A.P., Hanson, S.J., Klenk, H.-P., LaButti, K.M., Lapidus, A., Lindquist, E.A., Lipzen, A.M., Meier-Kolthoff, J.P., Ohm, R.A., Otillar, R.P., Pangilinan, J.L., Peng, Y., Rokas, A., Rosa, C.A., Scheuner, C., Sibirny, A.A., Slot, J.C., Stielow, J.B., Sun, H., Kurtzman, C.P., Blackwell, M., Grigoriev, I.V., Jeffries, T.W. (2016) Comparative genomics of biotechnologically important yeasts. Proceedings of the National Academy of Sciences 113(35), 9882–7, Doi: 10.1073/pnas.1603941113.Abstract/FREE Full Text

[58].Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M.A., Huelsenbeck, J.P. (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Systematic Biology 61(3), 539–42, Doi: 10.1093/sysbio/sys029.

[59].Rosewich, U.L., Kistler, H.C. (2000) Role of Horizontal Gene Transfer in the Evolution of Fungi. Annu. Rev. Phytopathol. 38(1), 325–63, Doi: 10.1146/annurev.phyto.38.1.325.

[60].Rychener, L., Albon, S.I., Djordjevic, S.P., Chowdhury, P.R., Ziech, R.E., de Vargas, A.C., Frey, J., Falquet, L. (2017) Clostridium chauvoei, an evolutionary dead-end pathogen. Frontiers in Microbiology 8(JUN), 1054, Doi: 10.3389/fmicb.2017.01054.

[61].Sb, H., T, D., Mk, W., F, C. (2020) Modulation of Peptidoglycan Synthesis by Recycled Cell Wall Tetrapeptides. Cell Reports 31(4), Doi: 10.1016/J.CELREP.2020.107578.

[62].Shaw, A.J., Hogsett, D.A., Lynd, L.R. (2009) Identification of the [FeFe]-

hydrogenase responsible for hydrogen generation in Thermoanaerobacterium saccharolyticum and demonstration of increased ethanol yield via hydrogenase knockout. Journal of Bacteriology 191(20), 6457–64, Doi: 10.1128/JB.00497-09.Abstract/FREE Full Text

[63].Shaw, A.J., Podkaminer, K.K., Desai, S.G., Bardsley, J.S., Rogers, S.R., Thorne, P.G., Hogsett, D.A., Lynd, L.R. (2008) Metabolic engineering of a thermophilic bacterium to produce ethanol at high yield. Proceedings of the National Academy of Sciences 105(37), 13769–74, Doi: 10.1073/pnas.0801266105.Abstract/FREE Full Text

[64].Smith, T.J., Hill, K.K., Foley, B.T., Detter, J.C., Munk, A.C., Bruce, D.C., Doggett, N.A., Smith, L.A., Marks, J.D., Xie, G., Brettin, T.S. (2007) Analysis of the neurotoxin complex genes in Clostridium botulinum A1-A4 and B1 strains: BoNT/A3, /Ba4 and /B1 clusters are located within plasmids. PLoS ONE 2(12), Doi: 10.1371/journal.pone.0001271.

[65].Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P., Jensen, L.J., Von Mering, C. (2019) STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Research 47(D1), D607–13, Doi: 10.1093/nar/gky1131.

[66].Tan, M.-F., Gao, T., Liu, W.-Q., Zhang, C.-Y., Yang, X., Zhu, J.-W., Teng, M.-Y., Li, L., Zhou, R. (2015) MsmK, an ATPase, Contributes to Utilization of Multiple Carbohydrates and Host Colonization of Streptococcus suis. PLOS ONE 10(7), e0130792, Doi: 10.1371/journal.pone.0130792.

[67].Tanaka, K.J., Song, S., Mason, K., Pinkett, H.W. (2018) Selective substrate uptake: The role of ATP-binding cassette (ABC) importers in pathogenesis. Biochimica et Biophysica Acta (BBA) -Biomembranes 1860(4), 868–77, Doi: 10.1016/j.bbamem.2017.08.011.

[68].Tian, L., Lo, J., Shao, X., Zheng, T., Olson, D.G., Lynd, L.R. (2016) Ferredoxin: NAD+ oxidoreductase of Thermoanaerobacterium saccharolyticum and its role in ethanol formation. Applied and Environmental Microbiology 82(24), 7134–41, Doi: 10.1128/AEM.02130-16.Abstract/FREE Full Text

[69].Törönen, P., Medlar, A., Holm, L. (2018) PANNZER2: A rapid functional annotation web server. Nucleic Acids Research 46(W1), W84–8, Doi: 10.1093/nar/gky350.

[70].Treangen, T.J., Rocha, E.P.C. (2011) Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. PLoS Genetics 7(1), e1001284, Doi: 10.1371/journal.pgen.1001284.

[71].Trifinopoulos, J., Nguyen, L.-T., von Haeseler, A., Minh, B.Q. (2016) W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. Nucleic Acids Research 44(18), 1–4, Doi: 10.1093/nar/gkw256.

[72].Tsakraklides, V., Shaw, A.J., Miller, B.B., Hogsett, D.A., Herring, C.D. (2012) Carbon catabolite repression in Thermoanaerobacterium saccharolyticum. Biotechnology for Biofuels 5(1), 85, Doi: 10.1186/1754-6834-5-85.

[73].Wang, W., Zhou, H., Ma, B., Owiti, A., Korban, S.S., Han, Y. (2016) Divergent Evolutionary Pattern of Sugar Transporter Genes is Associated with the Difference in Sugar Accumulation between Grasses and Eudicots. Sci Rep 6(1), 29153, Doi: 10.1038/srep29153.

[74].Webb, A.J., Homer, K.A., Hosie, A.H.F. (2008) Two closely related ABC transporters in Streptococcus mutans are involved in disaccharide and/or oligosaccharide uptake. Journal of Bacteriology 190(1), 168–78, Doi: 10.1128/JB.01509-07.Abstract/FREE Full Text

[75].Wilkens, S. (2015) Structure and mechanism of ABC transporters. F1000Prime Reports 7, Doi: 10.12703/P7-14.

[76].Williams, O.B., Morrow, M.B. (1928) The bacterial destruction of acetylmethyl-carbinol. Journal of Bacteriology 16(1), 43–8, Doi: 10.1128/jb.16.1.43-48.1928.FREE Full Text

[77].Wohlbach, D.J., Kuo, A., Sato, T.K., Potts, K.M., Salamov, A.A., Labutti, K.M., Sun, H., Clum, A., Pangilinan, J.L., Lindquist, E.A., Lucas, S., Lapidus, A., Jin, M., Gunawan, C., Balan, V., Dale, B.E., Jeffries, T.W., Zinkel, R., Barry, K.W., Grigoriev, I.V., Gasch, A.P. (2011) Comparative genomics of xylose-fermenting fungi for enhanced biofuel production. Proceedings of the National Academy of Sciences of the United States of America 108(32), 13212–7, Doi: 10.1073/pnas.1103039108.Abstract/FREE Full Text

[78].Xiao, Z., Xu, P. (2007) Acetoin metabolism in bacteria. Critical Reviews in Microbiology 33(2), 127–40, Doi: 10.1080/10408410701364604.

[79].Xu, X., Zeng, W., Li, Z., Wang, Z., Luo, Z., Li, J., Li, X., Yang, J. (2022) Genome-wide identification and expression profiling of sugar transporter genes in tobacco. Gene 835, 146652, Doi: 10.1016/j.gene.2022.146652.

[80].Yang, Z. (2007) PAML 4: Phylogenetic Analysis by Maximum Likelihood. Molecular Biology and Evolution 24(8), 1586–91, Doi: 10.1093/molbev/msm088.

[81].Zhang, J., Nielsen, R., Yang, Z. (2005) Evaluation of an Improved Branch-Site Likelihood Method for Detecting Positive Selection at the Molecular Level. Molecular Biology and Evolution 22(12), 2472–9, Doi: 10.1093/molbev/msi237.

[82].Zhao, B., Mesbah, N.M., Dalin, E., Goodwin, L., Nolan, M., Pitluck, S., Chertkov, O., Brettin, T.S., Han, J., Larimer, F.W., Land, M.L., Hauser, L., Kyrpides, N., Wiegel, J. (2011) Complete genome sequence of the anaerobic, halophilic alkalithermophile Natranaerobius thermophilus JW/NM-WN-LF. Journal of Bacteriology 193(15), 4023–4,

Doi: 10.1128/JB.05157-11.Abstract/FREE Full Text

[83].Zhao, Z., Xian, M., Liu, M., Zhao, G. (2020) Biochemical routes for uptake and conversion of xylose by microorganisms. Biotechnology for Biofuels 13(1), 21, Doi: 10.1186/s13068-020-1662-x.

[84].Zheng, T., Lanahan, A.A., Lynd, L.R., Olson, D.G. (2018) The redox-sensing protein Rex modulates ethanol production in Thermoanaerobacterium saccharolyticum. PLOS ONE 13(4), e0195143, Doi: 10.1371/journal.pone.0195143.

[85].Zhou, J., Olson, D.G., Lanahan, A.A., Tian, L., Murphy, S.J.L., Lo, J., Lynd, L.R. (2015) Physiological roles of pyruvate ferredoxin oxidoreductase and pyruvate formate-lyase in Thermoanaerobacterium saccharolyticum JW/SL-YS485. Biotechnology for Biofuels 8(1), 138, Doi: 10.1186/s13068-015-0304-1.

Discussion

Comparative Genomics and Machine Learning although already established for different biological studies, are incipient methodologies for understanding biological industrial applications. By combining these methodologies, researchers can gain a deeper understanding of the biology of microorganisms and design new biotechnological processes that are more efficient and environmentally friendly, with lower research and testing costs. The work on chapter 1 presents a classification model tailored to specifically identify transporters' ability to transport xylose. Using oversampling techniques and raising the prediction threshold, a reliable model was established which identified 25 potential xylose transporters, four of which were experimentally validated and named as SuL, SpX, SpH and SpG. All four transporters were capable of transporting xylose, highlighting the success of the model in prediction of this trait of interest. To create the model, thousands of features were extracted, either via the protr library, or through own calculations, from all the analyzed sequences, however a dimensionality reduction was needed to simplify predictions, reduce overfitting, and facilitate understanding of the most important parameters. This step was done through feature selection, and a few interesting features were returned which deserve to be commented.

The first interesting feature is an HMM extracted from the amino acid patterns in the transporters' pore, the most important amino acids for sugar affinity [24, 61]. Another group of features, similarly related, but extracted via protr were some PSSM identities and AA index, which have also been reported on literature as important for prediction of membrane transport sequences [50]. These features, coupled with the calculated HMM feature, show that there is an important hidden motif in the pore region that must be explored further to find the key amino acids for which rational engineering efforts can focus on.

A fourth feature that was returned from feature selection was the proportion of GFV tripeptides throughout the sequence. These tripeptides were all located on the transmembrane portions of the transporters, and it is still unclear their role with xylose transport, even though all predicted xylose transporters had a number of these motifs ranging from 1 to 3 in their sequence. Further studies should elucidate if these regions have a role with transporter stabilization or some other function related to sugar transport.

This evolutionary and sequence pattern overlap can also be seen on the probable convergent evolution of the site found under positive selection between the 3 *Spathaspora* chosen candidates and the *Sugiyamaella* transporter, as *Sugiyamaella* diverged much earlier but still has the same adaptations as the *Spathaspora* transporters, which corroborates

the utility of studying past adaptations with present sequence patterns.

The work on chapter 2 yielded noteworthy results for the studied xylose transporter families. Firstly, the known *E. coli* xylose MFS transporter xylE was placed in a family without any *T. sac* genes. No *T. sac* or other *Thermoanaerobacterium* genes were found to resemble xylE, with the closest match being a drug resistance transporter. This initial discovery indicates that *Thermoanaerobacterium*'s approach to xylose transport differs from other bacterial groups, potentially indicating a reason for their effectiveness in 2G co-fermentation. Secondly, no evidence for positive selection was found in the *T. sac* MFS transporters, which according to literature is expected. Interestingly enough, there weren't many duplicate copies, which is a common evolutionary strategy seen for transporters in yeast [45], but in prokaryotes it is known that gene duplication is not a main evolutionary driver [68]. Third, for the ABC transporters, evidence for positive selection was found for the xylose-specific extracellular binding protein (similar to xylF in *E. coli*) close to the predicted binding site. Due to the necessity of energy expenditure for ABC transporters, it is important that these extracellular binding proteins are extremely specific to the required sugar, so as not to waste energy, decreasing fitness. Signs of positive selection and increased expression during xylose fermentation, as shown in published microarray data [19], suggest this organism has an enhanced adaptation for xylose metabolism. This may improve its ability to capture xylose and contribute to its capability of co-fermenting glucose and xylose. Few xylose-specific sugar transporters are known and having a separate pathway for assimilation could alleviate competition for the active site of non-specific MFS transporters.

With the interesting evolutionary findings in the *T. sac* ABC transporters, it might be useful in next studies to create new machine learning models that try and better understand the features of these sequences, as a compliment to the MFS features, and verify if there are convergent patterns in the sugar binding strategies.

Overall it was shown that there are great benefits of prospecting novel genes of interest for industrial applications, especially sugar transporters, using evolutionary approaches. Both chapters presented in this work contributed with interesting new xylose transporters, some that have been further characterized experimentally, and others that can be further explored on future studies. For chapter 1, in addition to the applied knowledge, deep evolutionary studies have and are being conducted on the 182 yeasts data, contributing to the overall evolutionary knowledge of the Saccharomycotina clade. Coupled with algorithms that can detect present sequence patterns, a significant step was given to better understand the underlying mechanisms related to the uptake capacity of xylose by a given transporter. Together these strategies can help immensely during transporter prospection, by facilitating in silico search and finding proteins with higher chances of positive results in wet-lab validation, reducing time and costs associated with

research of this nature. As more data becomes available for the model, the better its ability to classify xylose tranporters will become, facilitating even further prospection. For chapter 2 the evolutionary resources for the subset of Firmicutes is available for further exploration outside of what has been discussed for xylose fermentation, with the additional provocation of a different strategy for xylose uptake via active transport, that can be explored for potential takeaways in yeast research.

Conclusions

2G ethanol can be a viable strategy to decrease dependency on fossil fuels while maintaining energy security with the possibility to integrate current established industries with these newer approaches, creating in the process new economic chains and incorporating ostracized regions into the economy. For these things to make sense, current issues with 2G ethanol must be solved, with the consumption of xylose by industrial microorganisms, mainly yeast, as an important point in this process. Xylose is carried to the cells through proteins called sugar transporters, which in yeast are of the MFS group. Unfortunately, the known xylose transporters are not efficient, and preferentially transport glucose in co-fermentation situations. This work attempts to address this issue by studying through comparative genomics and machine learning approaches what are the molecular and evolutionary keys that confers the ability to one transporter to carry xylose while another one has no such ability. In chapter 1 this has been explored by creating a machine learning model based on available data of known xylose transporters, which showed many interesting features, such as HMM of the pocket region and PSSMs of various protein regions, in accordance with the literature, where HMM and PSSM profiles are shown to be important in defining substrate specificity, indicating that xylose transport can in part be explained by amino acid patterns [24]. This model has been successfully used to select interesting transporters for wet-lab validation from a 182 yeast genomes dataset being used for comparative genomics study. These candidates showed promise for industrial applications as all of them were capable of transporting xylose, while the current and next iterations of the model showed promise as an initial screening for novel pentose transporters discovery. In chapter 2, an interesting organism, *T.sac*, was analysed also through comparative genomics to understand its general adaptations to xylose consumption, and specifically on sugar transporters, as it is able to co-ferment xylose and glucose. Positive selection in a specific xylose-binding protein of an ABC transporter was found. Studying more deeply the meaning of these adaptations may be interesting for rational engineering of known yeast transporters, especially the residues associated to the binding pocket. While at the moment replacing yeast as the main organism for industrial 2G production might not be the most optimal approach, understanding these adaptations in other species may bring insights and novelties for rational engineering of yeast or even change the state of the art in the future for biofuels production.

Bibliography

[1] Aditiya, H. B., Mahlia, T. M. I., Chong, W. T., Nur, H., and Sebayang, A. H. Second generation bioethanol production: A critical review. *Renewable and Sustainable Energy Reviews 66* (2016), 631–653. publisher: Elsevier Citation Key: Aditiya2016 ISBN: 13640321.

[2] Adler, M., Anjum, M., Berg, O. G., Andersson, D. I., and Sandegren, L. High fitness costs and instability of gene duplications reduce rates of evolution of new genes by duplication-divergence mechanisms. *Molecular Biology and Evolution 31*, 6 (6 2014), 1526–1535. PMID: 24659815 publisher: Oxford University Press.

[3] Alballa, M., and Butler, G. Toot-t: discrimination of transport proteins from non-transport proteins. *BMC Bioinformatics 21*, 3 (4 2020), 25.

[4] Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., and Aljaaf, A. J. A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. In *Supervised and Unsupervised Learning for Data Science*, M. W. Berry, A. Mohamed, and B. W. Yap, Eds., Unsupervised and Semi-Supervised Learning. Springer International Publishing, Cham, 2020, pp. 3–21.

[5] Amorim, H. V., and Lopes, M. L. Ethanol production in a petroleum dependent world: the brazilian experience. *Sugar Journal 67*, 12 (2005), 11–14. publisher: Kriedt Enterprises Ltd.

[6] Argueso, J. L., Carazzolle, M. F., Mieczkowski, P. A., Duarte, F. M., Netto, O. V. C., Missawa, S. K., Galzerani, F., Costa, G. G. L., Vidal, R. O., Noronha, M. F., Dominska, M., Andrietta, M. G. S., Andrietta, S. R., Cunha, A. F., Gomes, L. H., Tavares, F. C. A., Alcarde, A. R., Dietrich, F. S., McCusker, J. H., Petes, T. D., and Pereira, G. A. G. Genome structure of a saccharomyces cerevisiae strain widely used in bioethanol production. *Genome Research 19*, 12 (12 2009), 2258–2270. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press publisher: Cold Spring Harbor Lab PMID: 19812109.

[7] Balat, M. Production of bioethanol from lignocellulosic materials via the biochemical pathway : A review. *Energy Conversion and Management 52*, 2 (2011), 858–875. publisher: Elsevier Ltd Citation Key: Balat2011.

[8] Basso, L. C., de Amorim, H. V., de Oliveira, A. J., and Lopes, M. L. Yeast selection for fuel ethanol production in brazil. *FEMS Yeast Research 8*, 7 (11 2008), 1155–1163.

[9] Borelli, G., Fiamenghi, M. B., Dos Santos, L. V., Carazzolle, M. F., Pereira, G. A. G., José, J., and Zhang, G. Positive selection evidence in xylose-related genes suggests methylglyoxal reductase as a target for the improvement of yeasts' fermentation in industry. *Genome Biology and Evolution 11*, 7 (7 2019), 1923–1938. PMID: 31070742 publisher: Narnia.

[10] Brink, D. P., Borgström, C., Persson, V. C., Osiro, K. O., Gorwa-Grauslund, M. F., Ofuji Osiro, K., and Gorwa-Grauslund, M. F. D-xylose sensing in saccharomyces cerevisiae: Insights from d-glucose signaling and native d-xylose utilizers. *International journal of molecular sciences 22*, 22 (11 2021), 12410. PMID: 34830296 publisher: Multidisciplinary Digital Publishing Institute.

[11] Brown, C. J., Todd, K. M., and Rosenzweig, R. F. Multiple duplications of yeast hexose transport genes in response to selection in a glucose-limited environment. *Molecular Biology and Evolution 15*, 8 (8 1998), 931–942.

[12] Bueno, J. G. R., Borelli, G., Corrêa, T. L. R., Fiamenghi, M. B., José, J., de Carvalho, M., de Oliveira, L. C., Pereira, G. A. G., and dos Santos, L. V. Novel xylose transporter cs4130 expands the sugar uptake repertoire in recombinant saccharomyces cerevisiae strains at high xylose concentrations. *Biotechnology for Biofuels 13*, 1 (12 2020), 145. publisher: BioMed Central.

[13] Camargo, A. P., Sourkov, V., Pereira, G., and Carazzolle, M. Rnasamba: neural network-based assessment of the protein-coding potential of rna sequences. *NAR Genomics and Bioinformatics 2*, 1 (3 2020), lqz024.

[14] Cazenave, A., and Cozannet, G. L. Sea level rise and its coastal impacts. *Earth's Future 2* (2013), 15–34. Citation Key: Cazenave2013 ISBN: 23284277.

[15] Choi, C. C., and Ford, R. C. Atp binding cassette importers in eukaryotic organisms. *Biological Reviews 96*, 4 (8 2021), 1318–1330. publisher: John Wiley Sons, Ltd.

[16] Conant, G. C. Comparative genomics as a time machine: How relative gene dosage and metabolic requirements shaped the time-dependent resolution of yeast polyploidy. *Molecular Biology and Evolution 31*, 12 (12 2014), 3184–3193. PMID: 25158798 publisher: Oxford Academic.

[17] Cui, J., and Davidson, A. L. Abc solute importers in bacteria. *Essays in Biochemistry 50*, 1 (2011), 85–99. PMID: 21967053.

[18] Currie, D. H., Guss, A. M., Herring, C. D., Giannone, R. J., Johnson, C. M., Lankford, P. K., Brown, S. D., Hettich, R. L., and Lynd, L. R. Profile of secreted hydrolases, associated proteins, and slpa in thermoanaerobacterium saccharolyticum during the degradation of hemicellulose. *Applied and Environmental Microbiology 80*, 16 (2014), 5001–5011. PMID: 24907337 Citation Key: Currie2014.

[19] Currie, D. H., Raman, B., Gowen, C. M., Tschaplinski, T. J., Land, M. L., Brown, S. D., Covalla, S. F., Klingeman, D. M., Yang, Z. K., Engle, N. L., Johnson, C. M., Rodriguez, M., Joe Shaw, A., Kenealy, W. R., Lynd, L. R., Fong, S. S., Mielenz, J. R., Davison, B. H., Hogsett, D. A., and Herring, C. D. Genome-scale resources for thermoanaerobacterium saccharolyticum. *BMC Systems Biology 9*, 1 (2015), 1–15. PMID: 26111937 publisher: BMC Systems Biology Citation Key: Currie2015 ISBN: 1752-0509 (Electronic)\r1752-0509 (Linking).

[20] de Valk, S. C., Bouwmeester, S. E., de Hulster, E., and Mans, R. Engineering proton-coupled hexose uptake in saccharomyces cerevisiae for improved ethanol yield. *Biotechnology for Biofuels and Bioproducts 2022 15:1 15*, 1 (5 2022), 1–16. publisher: BioMed Central.

[21] Donzella, L., Sousa, M., and Morrissey, J. Evolution and functional diversification of yeast sugar transporters. *Essays in Biochemistry* (3 2023), EBC20220233.

[22] Donzella, L., Varela, J. A., Sousa, M. J., and Morrissey, J. P. Identification of novel pentose transporters in kluyveromyces marxianus using a new screening platform. *FEMS Yeast Research 21*, 4 (6 2021), 26. PMID: 33890624 publisher: Oxford Academic.

[23] dos Santos, L. V., de Barros Grassi, M. C., Gallardo, J. C. M., Pirolla, R. A. S., Calderón, L. L., de Carvalho-Netto, O. V., Parreiras, L. S., Camargo, E. L. O., Drezza, A. L., Missawa, S. K., Teixeira, G. S., Lunardi, I., Bressiani, J., and Pereira, G. A. G. Second-generation ethanol: The need is becoming a reality. *Industrial Biotechnology 12*, 1 (2 2016), 40–57. Citation Key: DosSantos2016a.

[24] Farwick, A., Bruder, S., Schadeweg, V., Oreb, M., and Boles, E. Engineering of yeast hexose transporters to transport d-xylose without inhibition by d-glucose. *Proceedings of the National Academy of Sciences of the United States of America 111*, 14 (2014), 5159–64. PMID: 24706835 Citation Key: Farwick2014 ISBN: 0027-8424.

[25] Gancedo, J. M. Yeast carbon catabolite repression. *Microbiology and Molecular Biology Reviews 62*, 2 (6 1998), 334–361. publisher: American Society for Microbiology.

[26] George, A. M., Ed. *ABC Transporters - 40 Years on*. Springer International Publishing, Cham, 2016. DOI: 10.1007/978-3-319-23476-2.

[27] Gong, C. S., Chen, L. F., Flickinger, M. C., Chiang, L. C., and Tsao, G. T. Production of ethanol from d-xylose by using d-xylose isomerase and yeasts. *Applied and Environmental Microbiology 41*, 2 (2 1981), 430–436. PMID: 16345717 PMCID: PMC243711.

[28] Gonçalves, C., Coelho, M. A., Salema-Oom, M., and Gonçalves, P. Stepwise functional evolution in a fungal sugar transporter family. *Molecular Biology and Evolution 33*, 2 (2 2016), 352–366. PMID: 26474848 ISBN: 0737-4038.

[29] Grassi, M. C. B., and Pereira, G. A. G. Energy-cane and renovabio: Brazilian vectors to boost the development of biofuels. *Industrial Crops and Products 129* (3 2019), 201–205.

[30] Gromiha, M. M., and Yabuki, Y. Functional discrimination of membrane proteins using machine learning techniques. *BMC Bioinformatics 9* (2008), 1–8. PMID: 18312695 Citation Key: Gromiha2008.

[31] Iwasaki, Y., Ikemura, T., Wada, K., Wada, Y., and Abe, T. Comparative genomic analysis of the human genome and six bat genomes using unsupervised machine learning: Mb-level cpg and tfbs islands. *BMC Genomics 23*, 1 (7 2022), 497.

[32] Joe Shaw, A., Hogsett, D. A., and Lynd, L. R. Natural competence in thermoanaerobacter and thermoanaerobacterium species. *Applied and Environmental Microbiology 76*, 14 (7 2010), 4713–4719. PMID: 20472726 publisher: American Society for Microbiology.

[33] Joe Shaw, A., Jenney, F. E., Adams, M. W. W., and Lynd, L. R. End-product pathways in the xylose fermenting bacterium, thermoanaerobacterium saccharolyticum. *Enzyme and Microbial Technology 42*, 6 (2008), 453–458. Citation Key: JoeShaw2008 ISBN: 0141-0229.

[34] Kim, S. R., Ha, S.-J., Wei, N., Oh, E. J., and Jin, Y.-S. Simultaneous co-fermentation of mixed sugars: a promising strategy for producing cellulosic ethanol. *Trends in Biotechnology 30*, 5 (5 2012), 274–282.

[35] Kondrashov, F. A. Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proceedings of the Royal Society B: Biological Sciences 279*, 1749 (12 2012), 5048–5057. publisher: Royal Society.

[36] Konno, N., and Iwasaki, W. Machine learning enables prediction of metabolic system evolution in bacteria. *Science Advances 9*, 2 (1 2023), eadc9130.

[37] Kumar, P., Barrett, D. M., Delwiche, M. J., and Stroeve, P. Methods for pretreatment of lignocellulosic biomass for efficient hydrolysis and biofuel production.

Industrial Engineering Chemistry Research 48, 8 (4 2009), 3713–3729. publisher: American Chemical Society.

[38] KUYPER, M., HARHANGI, H., STAVE, A., WINKLER, A., JETTEN, M., DELAAT, W., DENRIDDER, J., OPDENCAMP, H., VANDIJKEN, J., AND PRONK, J. High-level functional expression of a fungal xylose isomerase: the key to efficient ethanolic fermentation of xylose by ? *FEMS Yeast Research 4*, 1 (10 2003), 69–78. PMID: 14554198 Citation Key: KUYPER2003 ISBN: 1567-1356 (Print)\r1567-1356 (Linking).

[39] LAZAR, Z., NEUVÉGLISE, C., ROSSIGNOL, T., DEVILLERS, H., MORIN, N., ROBAK, M., NICAUD, J.-M., AND CRUTZ-LE COQ, A.-M. Characterization of hexose transporters in yarrowia lipolytica reveals new groups of sugar porters involved in yeast growth. *Fungal Genetics and Biology 100* (3 2017), 1–12.

[40] LEANDRO, M. J., FONSECA, C., AND GONÇALVES, P. Hexose and pentose transport in ascomycetous yeasts: an overview. *FEMS Yeast Research 9*, 4 (6 2009), 511–525.

[41] LEANDRO, M. J., GONÇALVES, P., AND SPENCER-MARTINS, I. Two glucose/xylose transporter genes from the yeast Candida intermedia : First molecular characterization of a yeast xylose–H + symporter. *Biochemical Journal 395*, 3 (May 2006), 543–549.

[42] LEE, Y. E., JAIN, M. K., LEE, C., LOWE, S. E., AND ZEIKUS, J. G. Taxonomic distinction of saccharolytic thermophilic anaerobes. *International Journal of Systematic Bacteriology 43*, 1 (1 1993), 41–51. publisher: Microbiology Society.

[43] LI, H., SCHMITZ, O., AND ALPER, H. S. Enabling glucose/xylose co-transport in yeast through the directed evolution of a sugar transporter. *Applied Microbiology and Biotechnology 100*, 23 (2016), 10215–10223. PMID: 27730335 publisher: Applied Microbiology and Biotechnology Citation Key: Li2016.

[44] LIN, H. H., HAN, L. Y., CAI, C. Z., JI, Z. L., AND CHEN, Y. Z. Prediction of transporter family from protein sequence by support vector machine approach. *Proteins 62*, 1 (2006), 218–31. PMID: 16287089 Citation Key: Lin2006.

[45] LIN, Z., AND LI, W. H. Expansion of hexose transporter genes was associated with the evolution of aerobic fermentation in yeasts. *Molecular Biology and Evolution 28*, 1 (1 2011), 131–142. PMID: 20660490 publisher: Oxford Academic.

[46] LOPES, M. L., PAULILLO, S. C. D. L., GODOY, A., CHERUBIN, R. A., LORENZI, M. S., GIOMETTI, F. H. C., BERNARDINO, C. D., AMORIM NETO, D. H. B., AMORIM, D. H. V., DE AMORIM NETO, H. B., DE AMORIM, H. V., AMORIM NETO, D. H. B., AND AMORIM, D. H. V. Ethanol production in brazil: a bridge between

science and industry. *Brazilian Journal of Microbiology 47* (2016), 1–13. publisher: Sociedade Brasileira de Microbiologia Citation Key: Lopes2016.

[47] Macedo, I. C., Seabra, J. E. A., and Silva, J. E. A. R. Green house gases emissions in the production and use of ethanol from sugarcane in brazil: The 2005/2006 averages and a prediction for 2020. *Biomass and Bioenergy 32*, 7 (7 2008), 582–595.

[48] Malina, C., Yu, R., Björkeroth, J., Kerkhoven, E. J., and Nielsen, J. Adaptations in metabolism and protein translation give rise to the crabtree effect in yeast. *Proceedings of the National Academy of Sciences 118*, 51 (12 2021), e2112836118. Company: National Academy of Sciences Distributor: National Academy of Sciences Institution: National Academy of Sciences Label: National Academy of Sciences publisher: Proceedings of the National Academy of Sciences.

[49] Marcet-Houben, M., and Gabaldón, T. Beyond the Whole-Genome Duplication: Phylogenetic Evidence for an Ancient Interspecies Hybridization in the Baker's Yeast Lineage. *PLOS Biology 13*, 8 (Aug. 2015), e1002220.

[50] Mishra, N. K., Chang, J., and Zhao, P. X. Prediction of membrane transport proteins and their substrate specificities using primary sequence information. *PLoS ONE 9*, 6 (2014), 3–6. PMID: 24968309 arXiv: http://www.ncbi.nlm.nih.gov/pubmed/24968309 Citation Key: Mishra2014.

[51] Mitrovic, D., McComas, S. E., Alleva, C., Bonaccorsi, M., Drew, D., and Delemotte, L. Reconstructing the transport cycle in the sugar porter superfamily using coevolution-powered machine learning. page: 2022.09.24.509294 section: New Results.

[52] Mohr, A., and Raman, S. Lessons from first generation biofuels and implications for the sustainability appraisal of second generation biofuels. *Energy Policy 63* (12 2013), 114–122.

[53] Nass, L. L., Pereira, P. A. A., and Ellis, D. Biofuels in brazil: An overview. *Crop Science 47*, 6 (2007), 2228–2237.

[54] Palma, M., Goffeau, A., Spencer-Martins, I., and Baret, P. V. A phylogenetic analysis of the sugar porters in hemiascomycetous yeasts. *Microbial Physiology 12*, 3-4 (2007), 241–248.

[55] Pfeiffer, T., and Morley, A. An evolutionary perspective on the crabtree effect. *Frontiers in Molecular Biosciences 1* (10 2014). [Online; accessed 2022-10-10].

[56] Piotrowski, J., Zhang, Y., Sato, T., Ong, I., Keating, D., Bates, D., and Landick, R. Death by a thousand cuts: the challenges and diverse landscape of

lignocellulosic hydrolysate inhibitors. *Frontiers in Microbiology 5* (2014). [Online; accessed 2023-03-28].

[57] Potter, N. I. How brazil achieved energy independence and the lessons the united states should learn from brazil's experience. *Washington University Global Studies Law Review 7*, 2 (1 2008). number: 2 publisher: Washington University in St. Louis School of Law.

[58] Quang, D., Chen, Y., and Xie, X. Dann: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics (Oxford, England) 31*, 5 (3 2015), 761–763. PMID: 25338716 PMCID: PMC4341060.

[59] Quistgaard, E. M., Löw, C., Guettou, F., and Nordlund, P. Understanding transport by the major facilitator superfamily (mfs): Structures pave the way. *Nature Reviews Molecular Cell Biology 17*, 2 (2016), 123–132. PMID: 26758938 publisher: Nature Publishing Group Citation Key: Quistgaard2016 ISBN: 1471-0080 (Electronic)\r1471-0072 (Linking).

[60] Raya, F. T., Marone, M. P., Carvalho, L. M., Rabelo, S. C., de Paula, M. S., Campanari, M. F. Z., Freschi, L., Mayer, J. L. S., Silva, O. R. R. F., Mieczkowski, P., Carazzolle, M. F., and Pereira, G. A. G. Extreme physiology: Biomass and transcriptional profiling of three abandoned agave cultivars. *Industrial Crops and Products 172* (11 2021), 114043.

[61] Reider Apel, A., Ouellet, M., Szmidt-Middleton, H., Keasling, J. D., and Mukhopadhyay, A. Evolved hexose transporter enhances xylose uptake and glucose/xylose co-utilization in saccharomyces cerevisiae. *Scientific Reports 6*, 1 (2016), 19512. PMID: 26781725 publisher: Nature Publishing Group Citation Key: ReiderApel2016 ISBN: 2045-2322 (Electronic) 2045-2322 (Linking).

[62] Riley, R., Haridas, S., Wolfe, K. H., Lopes, M. R., Hittinger, C. T., Göker, M., Salamov, A. A., Wisecaver, J. H., Long, T. M., Calvey, C. H., Aerts, A. L., Barry, K. W., Choi, C., Clum, A., Coughlan, A. Y., Deshpande, S., Douglass, A. P., Hanson, S. J., Klenk, H.-P., LaButti, K. M., Lapidus, A., Lindquist, E. A., Lipzen, A. M., Meier-Kolthoff, J. P., Ohm, R. A., Otillar, R. P., Pangilinan, J. L., Peng, Y., Rokas, A., Rosa, C. A., Scheuner, C., Sibirny, A. A., Slot, J. C., Stielow, J. B., Sun, H., Kurtzman, C. P., Blackwell, M., Grigoriev, I. V., and Jeffries, T. W. Comparative genomics of biotechnologically important yeasts. *Proceedings of the National Academy of Sciences 113*, 35 (2016), 9882–9887. PMID: 27535936 Citation Key: Riley2016.

[63] Sato, T. K., Tremaine, M., Parreiras, L. S., Hebert, A. S., Myers, K. S., Higbee, A. J., Sardi, M., McIlwain, S. J., Ong, I. M., Breuer, R. J.,

Avanasi Narasimhan, R., McGee, M. A., Dickinson, Q., La Reau, A., Xie, D., Tian, M., Reed, J. L., Zhang, Y., Coon, J. J., Hittinger, C. T., Gasch, A. P., and Landick, R. Directed evolution reveals unexpected epistatic interactions that alter metabolic regulation and enable anaerobic xylose use by saccharomyces cerevisiae. *PLOS Genetics 12*, 10 (10 2016), e1006372. Citation Key: Sato2016.

[64] Shaw, A. J., Podkaminer, K. K., Desai, S. G., Bardsley, J. S., Rogers, S. R., Thorne, P. G., Hogsett, D. A., and Lynd, L. R. Metabolic engineering of a thermophilic bacterium to produce ethanol at high yield. *Proceedings of the National Academy of Sciences 105*, 37 (2008), 13769–13774. PMID: 18779592 Citation Key: Shaw2008 ISBN: 1091-6490 (Electronic)\r0027-8424 (Linking).

[65] Stambuk, B. U., Dunn, B., Alves, S. L., Duval, E. H., and Sherlock, G. Industrial fuel ethanol yeasts contain adaptive copy number changes in genes involved in vitamin b1 and b6 biosynthesis. *Genome Research 19*, 12 (12 2009), 2271–2278. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press publisher: Cold Spring Harbor Lab PMID: 19897511.

[66] Subtil, T., and Boles, E. Competition between pentoses and glucose during uptake and catabolism in recombinant saccharomyces cerevisiae. *Biotechnology for Biofuels 5*, 1 (3 2012), 14. publisher: BioMed Central.

[67] Sànchez Nogué, V., and Karhumaa, K. Xylose fermentation as a challenge for commercialization of lignocellulosic fuels and chemicals. *Biotechnology Letters 37*, 4 (4 2015), 761–772.

[68] Treangen, T. J., and Rocha, E. P. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genetics 7*, 1 (1 2011), e1001284. PMID: 21298028 publisher: Public Library of Science.

[69] van Maris, A. J. A., Winkler, A. A., Kuyper, M., de Laat, W. T. A. M., van Dijken, J. P., and Pronk, J. T. Development of efficient xylose fermentation in saccharomyces cerevisiae: Xylose isomerase as a key component. In *Advances in biochemical engineering/biotechnology*, vol. 108. 2007, pp. 179–204. PMID: 17846724 DOI: $10.1007/10_2007_057 CitationKey : VanMaris2007 issue : AprilISSN : 0724-6145$.

[70] Wagner, A. Energy constraints on the evolution of gene expression. *Molecular Biology and Evolution 22*, 6 (6 2005), 1365–1374.

[71] Wang, M., Yu, C., and Zhao, H. Directed evolution of xylose specific transporters to facilitate glucose-xylose co-utilization. *Biotechnology and Bioengineering 113*, 3 (3 2016), 484–491. Citation Key: Wang2016.

[72] Wohlbach, D. J., Kuo, A., Sato, T. K., Potts, K. M., Salamov, A. A., Labutti, K. M., Sun, H., Clum, A., Pangilinan, J. L., Lindquist, E. A., Lucas, S., Lapidus, A., Jin, M., Gunawan, C., Balan, V., Dale, B. E., Jeffries, T. W., Zinkel, R., Barry, K. W., Grigoriev, I. V., and Gasch, A. P. Comparative genomics of xylose-fermenting fungi for enhanced biofuel production. *Proceedings of the National Academy of Sciences of the United States of America 108*, 32 (2011), 13212–7. PMID: 21788494 Citation Key: Wohlbach2011 ISBN: 1091-6490 (Electronic)\n0027-8424 (Linking).

[73] Wolfe, K. H. Origin of the Yeast Whole-Genome Duplication. *PLOS Biology 13*, 8 (Aug. 2015), e1002221.

[74] Zhang, X. C., Zhao, Y., Heng, J., and Jiang, D. Energy coupling mechanisms of mfs transporters. *Protein Science 24*, 10 (10 2015), 1560–1579. publisher: John Wiley Sons, Ltd.

[75] Zhou, H., Cheng, J.-s., Wang, B. L., Fink, G. R., and Stephanopoulos, G. Xylose isomerase overexpression along with engineering of the pentose phosphate pathway and evolutionary engineering enable rapid xylose utilization and ethanol production by saccharomyces cerevisiae. *Metabolic Engineering 14*, 6 (11 2012), 611–622.

[76] Zhu, Y., Zhang, J., Zhu, L., Jia, Z., Li, Q., Xiao, W., and Cao, L. Minimize the Xylitol Production in Saccharomyces cerevisiae by Balancing the Xylose Redox Metabolic Pathway. *Frontiers in Bioengineering and Biotechnology 9* (2021).

Appendix

APPENDIX A – Supplementary Files for Chapter 1

Table S1. Information of the species used for comparative genomics, including xylose fermentation or consumption capacity, and associated publication describing the phenotype.

Table S2. Xylose transporters used as positive samples for machine learning, including Uniprot accession, organism and publication describing xylose utilization.

Table S3. Primers used to validate transformation of EBY_Xyl1.

File S1. S. cerevisiae Codon optimized sequences for the four transporters chosen for characterization in xylose.

File S2. FASTA files of the four transporter families from the comparative genomics analysis.

Figure S1. BUSCO results for gene prediction.

Figure S2. UMAP spatial distribution of data after oversampling.

APPENDIX B – Supplementary Files for Chapter 2

Supplementary Figures.

Code repository for data used in this work

APPENDIX C – Other works

In addition to chapters 1 and 2 presented in this thesis, I have collaborated in many other publications within and outside my research group. Here are the works in which I participated or are participating, which are more relevant to the topics of this thesis:

1) **Positive Selection Evidence in Xylose-Related Genes Suggests Methylglyoxal Reductase as a Target for the Improvement of Yeasts' Fermentation in Industry**

https://doi.org/10.1093/gbe/evz036

This work was my first interaction with Comparative Genomics, in which we analysed 18 yeast genomes searching for adaptations possibly related to xylose metabolism which we could then suggest as points of interest for industrial biotechnology and genetic engineering. I was responsible for analysis of several gene families of interest, doing multiple sequence alignments, phylogenetic inferences, selection analysis, as well as species' phylogeny, gene gain and loss analysis, etc. I also contributed by writing parts of the manuscript, mainly parts of the methodology, results, and discussion sections.

2) **Novel xylose transporter Cs4130 expands the sugar uptake repertoire in recombinant *Saccharomyces cerevisiae* strains at high xylose concentrations**

https://doi.org/10.1186/s13068-020-01782-0

In this work Comparative Genomics was used to prospect xylose transporter candidates in the same dataset of 18 yeast genomes of work number 1. Briefly, the choice of candidates for wet lab work was done by first inferring phylogenetic relationships between the best described xylose transporters in literature and the sugar transporters from the 18 genomes dataset. Some of the closest transporters from the dataset to the best described transporters were then chosen for wet lab validation. I was already working on the 18 yeast genomes dataset for work number 1 and contributed to this work with the whole phylogenomic pipeline, helping to choose the transporter candidates for the experimental part of the project, and by writing part of the manuscript, mainly methodology, results and discussion sections.

3) **Structural and biochemical insights of xylose MFS and SWEET transporters in microbial cell factories: challenges to lignocellulosic hydrolysates fermentation**

in press

This work is a review of current knowledge of MFS and SWEET transporters applied to industrial applications. It is currently in prep and I have contributed to the text on the sections related to the evolution and structure of MFS and SWEET, the use of evolutionary methods as a screening strategy for biotechnological applications, and the use of machine learning to detect sequence patterns and predict xylose transport capacity

Annex

ANNEX A – Copyright declaration

Declaração

As cópias de artigos de minha autoria ou de minha co-autoria, já publicados ou submetidos para publicação em revistas científicas ou anais de congressos sujeitos a arbitragem, que constam da minha Dissertação/Tese de Mestrado/Doutorado, intitulada **Caracterização evolutiva e funcional de transportadores de xilose**, não infringem os dispositivos da Lei n.° 9.610/98, nem o direito autoral de qualquer editora.

Campinas, 05 de junho de 2023

Assinatura : 
Nome do(a) autor(a): **Mateus Bernabe Fiamenghi**
RG n.° 53.455.415-5

Assinatura : 
Nome do(a) orientador(a): **Gonçalo Amarante Guimarães Pereira**
RG n.° 56.797.320-7

ANNEX B – Bioethics/Biosecurity

UNICAMP

I N F O R M A Ç Ã O

INFORMAMOS que o projeto **CIBio/IB No. 2011/03 - Genômica e Biotecnologia**, cujo pesquisador responsável é o Prof. Dr. Gonçalo Amarante Guimarães Pereira, sub-projeto "Caracterização evolutiva e funcional de transportadores de xilose", do pós-graduando **Mateus Bernabe Fiamenghi**, encontra-se devidamente aprovado e regularizado junto a CIBio/IB-UNICAMP e a CTNBio, conforme legislação vigente.

Cidade Universitária "Zeferino Vaz",
23 de maio de 2019.

Prof. Dr. JOSÉ LUIZ PROENÇA MÓDENA
Presidente da CIBio
Instituto de Biologia – UNICAMP

CIBio/IB-Unicamp
Comissão Interna de Biossegurança
Instituto de Biologia - Unicamp
Caixa Postal 6109 - 13083-970 Campinas SP
Tel.: (19) 3521-6359 – e-mail: comisib@unicamp.br

Ministério do Meio Ambiente
CONSELHO DE GESTÃO DO PATRIMÔNIO GENÉTICO

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

Comprovante de Cadastro de Acesso

Cadastro nº A63BBD2

A atividade de acesso ao Patrimônio Genético, nos termos abaixo resumida, foi cadastrada no SisGen, em atendimento ao previsto na Lei nº 13.123/2015 e seus regulamentos.

Número do cadastro:	**A63BBD2**
Usuário:	**UNICAMP**
CPF/CNPJ:	**46.068.425/0001-33**
Objeto do Acesso:	**Patrimônio Genético**
Finalidade do Acesso:	**Pesquisa**

Espécie

Saccharomyces cerevisiae

Spathaspora spp

Candida spp

Candida boidinii

Candida sojae

Título da Atividade: **LGE 01.21**

Equipe

Gonçalo Amarante Guimarães Pereira	**UNICAMP**
Monique Furlan	**Unicamp**
Juliana Pimentel Galhardo	**Unicamp**
Antônio José Rocha	**Unicamp**
Antônio Pedro de Castello Branco da Rocha Camarg	**Unicamp**

André Pfaffenbach Piffer	**Unicamp**
Luciana Souto Mofatto	**Unicamp**
Guilherme Borelli	**Unicamp**
Thaís Oliveira Secches	**Unicamp**
Duguay Rodrigues Monteiro da Silva	**Unicamp**
Milena Antunes Piccart Gutierrez	**Unicamp**
Alexandra Russolo Cardelli	**Unicamp**
Larissa Escalfi Tristão	**Unicamp**
Jade Ribeiro dos Santos	**Unicamp**
Beatriz de Oliveira Vargas	**Unicamp**
Marcelo Falsarella Carazzolle	**Unicamp**
Lucas Miguel de Carvalho	**Unicamp**
WELBE OLIVEIRA BRAGANÇA	**Unicamp**
ANDERSON FERREIRA DA CUNHA	**UFSCAR**
ANDRÉ RICARDO ALCARDE	**ESALQ - USP**
JULIANA VELASCO	**CNPEM**
MIRTA NATALIA COUTOUNÉ	**UNICAMP/CNPEM**
GABRIELA VAZ MEIRELLES	**CNPEM**
IRAN MALAVAZI	**UFSCAR**
JÚLIA ANALIA OLIVEIRA HANSEN	**UNICAMP**
ROSANA VASCO DAS CHAGAS	**UNICAMP**
Frank Uriel Suarez Lizarazo	**UNICAMP**
Fellipe da Silveira Bezerra de Mello	**UNICAMP**
Lethicia Camboin de Oliveira	**UNICAMP**
Mateus Bernabé Fiamenghi	**UNICAMP**

Parceiras Nacionais

45.358.058/0001-40 / FUNDACAO UNIVERSIDADE FEDERAL DE SAO CARLOS

63.025.530/0025-81 / Escola Superior de Agricultura Luiz de Queiroz

01.576.817/0001-75 / CENTRO NACIONAL DE PESQUISA EM ENERGIA E MATERIAIS

63.025.530/0028-24 / Escola de Engenharia de São Carlos – USP

13.808.281/0001-55 / BIOCELERE AGROINDUSTRIAL LTDA

Parceiras no Exterior

Duke University

University of North Carolina at Chapel Hill

Colorado State University

Data do Cadastro: **13/01/2021 15:21:59**

Situação do Cadastro: **Concluído**