Machine learning enables identification of a novel yeast galactose utilization pathway

Marie-Claire Harrison[1], Emily J. Ubbelohde[2], Abigail L. LaBella[1,3], Dana A. Opulente[2,4], John F. Wolters[2], Xiaofan Zhou[5], Xing-Xing Shen[6], Marizeth Groenewald[7], Chris Todd Hittinger[2,*], and Antonis Rokas[1,*]

1) Department of Biological Sciences and Evolutionary Studies Initiative, Vanderbilt University, Nashville, TN 37235, USA

2) Laboratory of Genetics, DOE Great Lakes Bioenergy Research Center, Center for Genomic Science Innovation, J. F. Crow Institute for the Study of Evolution, Wisconsin Energy Institute, University of Wisconsin-Madison, Madison, WI 53726, USA

3) Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, Charlotte, NC 28262, USA

4) Department of Biology, Villanova University, Villanova, PA 19085, USA

5) Guangdong Province Key Laboratory of Microbial Signals and Disease Control, Integrative Microbiology Research Center, South China Agricultural University, Guangzhou 510642, China

6) Key Laboratory of Biology of Crop Pathogens and Insects of Zhejiang Province, Institute of Insect Sciences, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou 310058, China

7) Westerdijk Fungal Biodiversity Institute, Utrecht 3584, The Netherlands

*Corresponding author: Chris Todd Hittinger and Antonis Rokas.

Email: cthittinger@wisc.edu and antonis.rokas@vanderbilt.edu

Author Contributions: M.C.H., E.J.U., C.T.H. and A.R. designed the research; M.C.H. and E.J.U. performed the research; M.C.H., E.J.U., A.L.L., D.A.O., J.F.W., X.Z., X.X.S., M.G., C.T.H. and A.R. contributed new reagents/analytic tools; M.C.H., E.J.U., A.L.L., D.A.O., C.T.H. and A.R. analyzed the data; and M.C.H., E.J.U, C.T.H. and A.R. wrote the paper.

Competing Interest Statement: A. R. is a scientific consultant for LifeMine Therapeutics, Inc. The authors declare no other competing interests.

Classification: Biological Sciences, Evolution

Keywords: *GAL* pathway, fungal evolution, primary metabolism, random forest, artificial intelligence, galactitol.

Abstract

How genomic differences contribute to phenotypic differences is a major question in biology. The recently characterized genomes, isolation environments, and qualitative patterns of growth on 122 sources and conditions of 1,154 strains from 1,049 fungal species (nearly all known) in the yeast subphylum Saccharomycotina provide a powerful, yet complex, dataset for addressing this question. We used a random forest algorithm trained on these genomic, metabolic, and environmental data to predict growth on several carbon sources with high accuracy. Known structural genes involved in assimilation of these sources and presence/absence patterns of growth in other sources were important features contributing to prediction accuracy. By further examining growth on galactose, we found that it can be predicted with high accuracy from either genomic (92.2%) or growth data (82.6%) but not from isolation environment data (65.6%). Prediction accuracy was even higher (93.3%) when we combined genomic and growth data. After the *GAL*actose utilization genes, the most important feature for predicting growth on galactose was growth on galactitol, raising the hypothesis that several species in two orders, Serinales and Pichiales (containing the emerging pathogen *Candida auris* and the genus *Ogataea*, respectively), have an alternative galactose utilization pathway because they lack the *GAL* genes. Growth and biochemical assays confirmed that several of these species utilize galactose through an oxidoreductive D-galactose pathway, rather than the canonical *GAL* pathway. Machine learning approaches are powerful for investigating the evolution of the yeast genotype-phenotype map, and their application will uncover novel biology, even in well-studied traits.

Significance Statement

Can we predict which organisms will grow on a particular type of sugar from genetic data or from knowing on what other sugars they can grow? To answer these and similar questions, we used an artificial intelligence algorithm on a one-of-a-kind dataset of genomic, metabolic, and ecological data from more than a thousand yeast species. The algorithm predicted organisms' growth on different sugars from their genomic data or from patterns of growth on other sugars with high accuracy. Focusing on galactose, a sugar found in milk and numerous plant products, our algorithm helped us discover yeast species that grow on galactose through a previously unknown metabolic pathway, illustrating the potential power of artificial intelligence in biological discovery from big data.

Main Text

Introduction

Yeasts in the subphylum Saccharomycotina (hereafter referred to as yeasts) are genomically diverse, geographically widely distributed, found in diverse habitats, and utilized for diverse purposes by humans – the baker's yeast *Saccharomyces cerevisiae* is the cornerstone of the winemaking, brewing, baking, and biotech industries; *Candida albicans* is a human commensal that thrives in the human gut and occasionally becomes a serious pathogen; *Candida auris* is an emerging fungal pathogen of great concern because of its innate resistance to available antifungal drugs; and *Lipomyces starkeyi* prodigiously produces lipids and has several biotechnological applications (1–3) .

Yeast ecological diversity is thought to be intimately tied to the vast diversity in their diets, i.e., the diversity of primary metabolic capabilities that allow them to grow on many different sources of carbon and nitrogen (4). However, we currently lack a comprehensive understanding of how variation in yeast gene content or regulation is related to the metabolic diversity and environmental adaptation of the ~1,200 species found across the subphylum. Recently, the Y1000+ Project (http://y1000plus.org/) published draft genome sequences of 1,154 representative strains (mostly taxonomic type strains) from 990 described and 61 candidates for

new species of yeasts (5–7). The Y1000+ Project has also systematically recorded (from the literature) and/or experimentally generated the isolation environments and qualitative and quantitative patterns of growth on diverse carbon sources, nitrogen sources, and environmental conditions (e.g., temperature and salinity) for a very large fraction of the same set of strains (4, 6). The availability of a comprehensive dataset that captures the vast genomic, environmental, and metabolic diversity of yeasts provides a unique testbed for understanding how adaptation to unique environments occurs in eukaryotic genomes (7).

Several of the pathways that allow yeasts to grow on certain sources are well-characterized(8) (Riley et al 2016). For example, sucrose assimilation depends on the invertase Suc2p, and maltose assimilation depends on the maltose permease Mal31p and maltase (α-D-glucosidase) Mal32, which can also act on sucrose (9, 10). Arguably the best studied pathway is the Leloir or *GAL*actose utilization pathway, which has become a model not only for understanding gene regulation in eukaryotes (11, 12), but also for how evolutionary changes in gene sequences, arrangement, and regulation contribute to ecological adaptation (13–19). In the *GAL* pathway of the baker's yeast *Saccharomyces cerevisiae*, Gal2p or an Hxt transporter protein imports D-galactose into the cell, where the mutarotase domain of Gal10p acts on the sugar, if necessary. Then, Gal1p converts it to galactose-1-phosphate, representing the first energy-consuming step of the pathway (20). Gal7p then converts galactose-1-phosphate to UDP-galactose. Gal10p acts on UDP-galactose using its epimerase domain, resulting in the production of UDP-glucose. Finally, Gal7p converts UDP-glucose to glucose-1-phosphate, which Pgm1p/Pgm2p then converts to glucose-6-phosphate, which enters glycolysis to produce energy for the cell (20).

Galactose abundance varies widely across yeast environments. For example, due to both the dietary influx of galactose and the synthesis of the sugar, galactose is abundant in the gut, bloodstream, and urine of most mammals (including humans) in the form of oligosaccharides, glycoproteins, and glycolipids, as well as in milk and other dairy products in the form of lactose (a disaccharide composed of galactose and glucose subunits) (21). Galactose is also found in a variety of fruit, vegetable, and other plant products, such as legumes; levels of galactose in common fruits and vegetables range from <0.1 mg/100 g to 34 mg/100 g (22–24). Galactose is also part of oligosaccharides, such as lactose, raffinose, and melibiose, as well as glycoproteins and glycolipids, that vary in their distribution across environments (23, 24); hydrolysis of these molecules by microbial enzymes can release free galactose.

The substantial variation in abundance of galactose in different environments is reflected in the evolution of the *GAL* pathway and its regulation across the subphylum Saccharomycotina. Numerous instances of wholesale pathway loss and gain, including by horizontal gene transfer, have been discovered (13, 15, 16, 19), as well as striking instances of ancient, multi-locus polymorphisms within species (16–18, 25). Different regulatory systems that lead to different modes of induction and rates of growth have also evolved in different lineages. For example, *C. albicans* exhibits an earlier graded induction in response to galactose, while *S. cerevisiae* has a more bimodal expression (14, 26, 27).

The rich genomic, environmental, and metabolic data of the Y1000+ Project, coupled with extensive genetic and biochemical knowledge of yeast primary metabolism, provide a unique opportunity to explore the genotype-phenotype map, which models the interaction between the genes and the traits of an organism, and how it has evolved across a subphylum. However, the enormity and complexity of the Y1000+ Project's data make standard statistical analyses less suitable. In recent years, machine learning algorithms have emerged as powerful tools for analyzing biological big data (28). Examples include predicting genes involved in specialized metabolism (29), predicting the bioactivities of specialized metabolites from genomic data (30, 31), predicting protein expression and function from regulatory and protein sequences (32–34),

and distinguishing fungal ecological lifestyles, such as saprobes from plant pathogens (35) or generalists from specialists (6).

One of the most successful machine learning algorithms for analyzing biological datasets is the random forest algorithm, which employs randomized decision trees trained on subsets of the data to identify the most informative data features (e.g., a gene's presence / absence, a gene's function, a strain's ability to grow on a given substrate) for predicting a trait of interest (e.g., the ability to assimilate galactose). The algorithm is known to perform well in biological datasets, likely because it can handle datasets where the number of variables is larger than that the number of observations (36), it can be trained on a part of the dataset at a time, and it can capture interactive effects between features (37). Identification of the most important features that contribute to the prediction accuracy of the random forest algorithm is straightforward and efficient, facilitating the exploration of very large datasets for biological meaning and the generation of testable hypotheses.

In this study, we used a random forest algorithm trained on environmental, metabolic, and/or genomic data to predict the growth of nearly all known species of Saccharomycotina on different carbon sources (Figure 1, Tables S1 – S4). Predicting growth on 29 different carbon sources tended to be highly accurate when the algorithm was trained on gene presence/absence and/or on presence/absence of growth on other carbon sources, which shows that both metabolic genes and the structure of the metabolic network are highly informative for understanding the evolution of yeast primary metabolism; in contrast, the predictive ability of isolation environment data was weak. Although the most important features associated with prediction accuracy were well-known genes and carbon sources associated with the source of interest, our machine learning approach also identified novel features not previously known to be associated with growth on a given carbon source. To illustrate the predictive ability of our approach, we used growth on galactose as a test case because our machine learning approach suggested a possible novel alternative pathway for galactose assimilation in the genus *Ogataea* and in a clade containing *C. auris*, which both lack *GAL* genes. Growth and biochemical assays validated that these species assimilate galactose through a hypothesized oxidoreductive D-galactose pathway, demonstrating the potential power of machine learning analysis for studying the relationship between genomic and phenotypic variation across vast evolutionary timescales.

Results

Machine learning accurately predicts growth on 29 different carbon sources from metabolic and genomic data but not from environmental data

A random forest algorithm (Figure 1) trained on the metabolic data matrix had high balanced accuracy (on average, 82%) for predicting growth of the 893 strains representing 885 of the Y1000+ yeast species on 29 different carbon sources. This result indicates that variation in the content and structure of the primary metabolic network in different strains informs patterns of growth on these substrates (Figure 2, Table S5). A random forest algorithm trained on the genomic data matrices (comprised of InterPro and/or KEGG Orthology (KO) annotations) was similarly accurate for predicting growth on these 29 sources (on average, 80-81% balanced accuracy). Interestingly, KO annotations were able to predict growth on substrates, such as D-xylose (~82% accurate) and L-sorbose (~79% accurate), with good accuracy; several previous studies have noted that the utilization of these substrates cannot be inferred solely from patterns of gene presence / absence, since the presence of certain genes (e.g., the *XYL* genes) is required for growth on these substrates but is not sufficient for predicting the ability to grow on them (8, 38–40).

In contrast, when the random forest algorithm was trained on environmental datasets, the balanced accuracy was between 49-60% (on average, 55%), which is only marginally above random accuracy (Figure 2, Table S5). This result suggests that our environmental dataset does not provide useful predictors for growth on these sources. Examination of the ROC/AUC curves, confusion matrices, and most important features for predicting growth on xylose, sucrose, and galactose supports this hypothesis: accuracy is only marginally above random using environmental data, and the most important features concern isolation environments not known to have high amounts of these sugars (Figure S1).

However, the accuracy of predicting growth on 29 carbon sources using a random forest algorithm trained on isolation environments was on average 60% when only specialists were included in the analysis, which compared favorably to 54% average accuracy when only generalists were included and 55% accuracy when all species were included. This result suggests that isolation environment is more informative for predicting carbon utilization of specialists (Figure S5, Table S12). Additionally, generalists tended to be better predicted on more commonly utilized substrates, while specialists were better predicted on more rarely utilized substrates (Figure S5, Table S12).

Top features for predicting growth on a specific carbon source are related sources and metabolic genes

The top features for predicting growth on the 29 carbon sources examined were often biologically relevant (Figure 2, Figure 3, Table S5). For example, for xylose, the most important feature was growth on xylitol, a metabolic intermediate in the typical xylose-degrading pathway in yeasts and other fungi (39, 41), while for sucrose, the most important feature was maltose, another disaccharide containing a glucose moiety (10) (Figure 3). For galactose, the top features included 2-keto-D-gluconate and L-sorbose, which are generated from glucose or galactose, respectively, by the enzymes acting on an alternative galactose-degrading pathways in some bacteria and fungi (41–44), as well as lactose and melibiose, disaccharides that contain galactose (Figure 3). When the top feature from each metabolic trait matrix was removed for xylose, sucrose, and galactose, and then the random forest was re-ran recursively, accuracy decreased rapidly at first for sucrose and more slowly for xylose and galactose, even though xylose and galactose were initially less accurate (~80% accuracy) than sucrose (~90% accurate) (Figure S4, Table S11). After removing the top feature from the algorithm around 30 times, the accuracy of predicting growth on xylose, sucrose, and galactose remained around 60%-70% and continued to slowly

decline (Figure S4, Table S11). At around 90 top features removed, accuracy started declining more steeply toward 50% or random accuracy (Figure S4, Table S11). This analysis demonstrates how much connectivity there is between metabolic traits in the random forest algorithm, as there remains a moderate level of accuracy even while removing top related traits to each carbon substrate.

A random forest algorithm trained on KEGG Orthology (KO) annotations was similarly accurate for predicting growth on xylose (~82%), sucrose (~87%), and galactose (~91%) to the combined KO and InterPro genomic dataset (Figure 2, Figure 3). Despite the larger size of the genomic data matrix (over 5,000 features compared to the metabolic data matrix of 122 features), the top features of the genomic data matrix were still often related to genetic pathways or enzymes known to be involved in the utilization of each source. The top features for the highly accurate prediction of growth on galactose were *GAL7* and *GAL10* (specifically the mutarotase domain), which are parts of the yeast *GAL* pathway (13). Despite the mis-annotation of the yeast *GAL1* by KO (see Methods), the algorithm was still nearly as accurate when trained on the entire genomic data matrix as when trained on the manually curated *GAL* gene orthologs (Figure 5). The top feature for the algorithm predicting growth on sucrose was oligo-1,6-glucosidase (K01182), which corresponds to the α-glucosidases encoded by *MAL32* and *MAL12*, as well as *IMA1-IMA5*, which indeed do act on sucrose, as well as maltose in some yeasts (9, 10). The distribution of *XYL1, XYL2,* and *XYL3* does not always correlate with yeast growth on xylose (8, 40). Even though the *XYL* genes were present in the KO database (except for *XYL3*, which was misannotated), they were not among the top features contributing to the 85% prediction accuracy, but an α-xylosidase (K01811) was the fifth most important feature (Figure 3). Since galactose metabolism and its associated genetic pathway has been thoroughly studied in yeasts, the remainder of this paper is focused on using growth on galactose as a test case for the utility of this machine-learning pipeline.

The* GAL *genes are highly predictive of growth on galactose in most, but not all, yeasts

Plotting the presence/absence of the *GAL* genes jointly with the presence/absence of growth on galactose on genome-scale phylogeny of 1,154 yeast strains showed that the distributions of the *GAL* genes were tightly correlated with the distribution of growth on galactose. Specifically, 526/558 strains that can grow on galactose have the *GAL* genes, and 277/310 strains that cannot grow on galactose lack the *GAL* genes. Notably, there are two lineages in the orders Serinales and Pichiales that can grow on galactose but lack the *GAL* genes (Figure 4). One lineage contains species closely related to the emerging opportunistic pathogen *Candida auris* in the order Serinales. The second lineage contains species belonging to the genus *Ogataea* in the order Pichiales. Isolation environments, such as isolation from plants, showed no significant association with growth on galactose (Figure 4).

Using the scores from the sequence similarity searches (from the jackhmmer software) of *GAL1, GAL7*, *GAL102,* and *GAL10*, the algorithm was even more accurate in its predictions of growth on galactose (92.2%). When the metabolic dataset was added to the training data, the accuracy increased even further to 93.1% (Figure 5). This increase in accuracy suggests that there are strains for which presence or absence of the *GAL* genes cannot accurately predict growth on galactose; if that were the case, then the increase in accuracy due to the inclusion of the rest of the metabolic dataset raises the possibility that there might be an alternative galactose-degrading pathway in some yeasts. After the *GAL* genes, the most predictive feature was growth on galactitol, pointing to a possible role for this metabolite as an intermediate in a potential alternative pathway (Figure 5). Previous work in filamentous fungi identified a galactose-

degrading pathway that involves galactitol as an intermediate(42, 45), leading us to hypothesize that a similar pathway may be present in these yeasts and contribute to the increase in accuracy.

Machine learning predicts an alternative galactose-degrading pathway in two yeast lineages that lack GAL *genes*

To further explore the possibility of an alternative galactose utilization pathway that uses galactitol as an intermediate, we trained our random forest algorithm just on the *GAL* genes and growth on galactitol. We found that this algorithm was almost as accurate as when the rest of the metabolic dataset was added (93.1% versus 93.3%). Examination of the confusion matrices when the algorithm was trained using just the *GAL* gene data versus when trained on the *GAL* gene data and metabolic data suggested that the increase in accuracy came from 16 species that were previously classified as false negatives and were now true positives (Figure 6). Since these species lack the *GAL* genes, our original algorithm predicted that they could not grow on galactose; when growth on galactitol was added, however, they were correctly predicted to grow on galactose, further supporting the hypothesis that they have an alternative galactose-degrading pathway (Figure 6). These 16 species are all able to grow on galactitol and belong to the two lineages that lack *GAL* genes, as noted previously in Figure 4: the lineage of species closely related to *Candida auris* in Serinales and the genus *Ogataea* in Pichiales. Even with this highly accurate algorithm, there were several species that were still not correctly predicted: 22 false negatives (strains that are predicted not to grow, but do) (Table S7) and 35 false positives (strains that are predicted to grow, but do not) remained, plus 3 species with low *GAL* gene sequence similarity scores also became false positives with this new algorithm, bringing the total to 38 false positives (Table S8). These species warrant further investigation as they may contain other alternative pathways, grow weakly on galactose or only under specific conditions (46), use galactose in glycosylation but not for assimilation (as the fission yeast *Schizosaccharomyces pombe*) (47), or have pseudogenized *GAL* genes (48). We note that the *GAL* genes of yeasts that were false positives in our classification exhibited, on average, lower sequence similarity scores in our *GAL* gene searches than the *GAL* genes of yeasts that were true positives (Table S9), which is consistent with reduced purifying selection.

Some Pichiales and Serinales species utilize galactose through an oxidoreductive galactose utilization pathway

To test the hypothesis that some species lacking *GAL* pathways can indeed utilize galactose, we tested three species (Table S10) from two different orders, *C. ruelliae* and *C. duobushaemulonii* from Serinales and *O. methanolica* from Pichiales (49), for growth on galactose as the sole carbon source and measured galactose consumption. All three species grew to high cell densities and accumulated more biomass than the *S. cerevisiae* positive control (Figure S3A), which contains an intact *GAL* pathway. Sugar quantification indicated galactose consumption in all three species (Figure 7A). The first step of the known oxidoreductive galactose pathway in species of *Aspergillus* fungi (outside of Saccharomycotina yeasts) utilizes an aldose reductase, which reduces galactose to the sugar alcohol galactitol while oxidizing NADPH to $NADP^+$ (50) (Figure 7B). Thus, we developed a biochemical assay for NADPH-dependent enzymatic activity on galactose as the sole carbon source. In this assay, species that exhibit the hypothesized enzymatic activity are predicted to show a decrease in NADPH absorbance at 340 nm over time, while species that do not exhibit enzymatic activity are predicted to show no decrease in NADPH over time (Figure 7C). All three species displayed decreases in absorbance of NADPH compared to their respective negative controls with no substrate (Figure 7D) and no extracted protein (Figure S3B), which indicates that the cells express NADPH-dependent enzymatic activity that is dependent on the presence of galactose. The *S. cerevisiae* negative control used for this experiment possessed an intact *GAL* pathway and did not show a decrease in NADPH absorbance over time, indicating a lack of NADPH-dependent enzymatic activity on galactose as

the sole carbon source. Thus, we conclude that these three species possess at least the first step of an oxidoreductive pathway.

Discussion

In this study, we employed machine learning on the rich environmental, metabolic, and genomic data from nearly all known species of an entire eukaryotic subphylum to predict patterns of yeast growth on different carbon sources. We found that we could accurately predict growth on diverse sources of carbon from genomic and/or metabolic data but not from environmental data (Figure 2). Previous research showed that many yeast traits are connected in a trait-trait network, likely due to shared genes in different metabolic pathways (4, 6). These connections and overlap in gene functions likely explain the high accuracy of prediction from metabolic and/or genomic data. Interestingly, accuracy of prediction was high, even for carbon sources for which enzyme specificity was lacking, such as xylose (Figure 3) (40). However, accuracy for xylose growth was lower than for predicting growth on sources, such as galactose, whose utilization pathways contain dedicated enzymes (Figure 3).

In contrast, the accuracy of prediction of growth on different carbon sources from isolation environment data was marginally better than random (Figure 3). There are two possible explanations for this finding. The first is that isolation environments may be heterogenous in their carbon sources and thus capable of supporting metabolically diverse yeast species. An alternative, not necessarily mutually exclusive explanation, is that isolation environments can be informative with respect to yeast diets, but that our current environmental data are incomplete. Notably, our isolation environmental data for each yeast included in the data matrix stem from information present in the taxonomic description of the type strain of each species. A dataset that contains the range of isolation environments of each yeast species would potentially be much more informative but is currently unavailable.

We also found that machine learning accuracy for predicting growth on galactose was higher when both the presence / absence of *GAL* genes and growth on galactitol were used in training compared to just the presence / absence of the *GAL* genes alone (Figure 5), suggesting the presence of a rare alternative galactose-degrading pathway. We discovered that this alternative galactose-degrading pathway is found in two distinct lineages that grow in galactose in the absence of *GAL* genes; we further proposed that this alternative pathway involves galactitol as a metabolic intermediate (Figures 4-6). Enzyme assays validated the oxidoreductive activity of three species in these two lineages when grown on galactose, providing additional support for the hypothesized mechanism of utilization (Figure 7). We are currently investigating which genes are involved in this alternative pathway.

This work illustrates the remarkable breadth of yeast metabolic diversity and how machine learning approaches can help uncover novel biology, even in well-studied traits, such as galactose assimilation. The potential for additional discoveries using machine learning is further highlighted by considering the several yeasts that appear as false positives or false negatives in our machine learning predictions. There are several possible explanations for why we currently cannot accurately predict growth on galactose for every strain in the subphylum. One explanation for some of the false positives could be that the *GAL* pathway is inactivated in some of the strains examined, but that their genomes contain *GAL* pseudogenes. Examples of *GAL* pseudogenes are known from several different species (16, 18, 48), but strains with pseudogenes would still give positive hits in our ortholog detection analyses. In support of this hypothesis, the average sequence similarity scores for the *GAL* genes in yeasts classified as false positives were lower than the scores for *GAL* genes in yeasts classified as true positives (Tables S8 and S9). Another possible explanation for false positives could be that some yeasts may contain *GAL* genes that are used in other processes, such as glycosylation, but not in assimilation; although such

examples are not currently known from the Saccharomycotina, the fission yeast *Schizosaccharomyces pombe* (subphylum Schizosaccharomycotina) is a case in point (51). They may also be growing very weakly or under specific conditions not tested here. Furthermore, since growth on galactitol is predictive of this alternative pathway of galactose utilization in the genus *Ogatea* and the *C. auris* lineage, our algorithm now predicts that any strain that grows on galactitol can also grow on galactose, which may not always true (e.g., some yeasts in these lineages may be lacking the gene(s) to convert galactose to galactitol). In fact, there are six yeasts (five from these two lineages, plus one *Starmerella* species) in the list of false positives that grow in galactitol but do not grow in galactose. Finally, we note that there are more false positives in lineages other than the more extensively studied Serinales and Saccharomycetales; this could be because the availability of fewer strains from other lineages results in less accurate identification of gene presence/absence. Alternatively, the induction of *GAL* genes or use of the pathway may be different in these lineages (Table S8).

Yeasts that appear as false negatives in our analyses, which indicates that they can indeed grow on galactose but are not predicted to grow, may be growing weakly or they may have other alternative pathways that do not involve galactitol. These may also lack the appropriate inducing conditions for growth on galactitol since they are often closely related to our documented alternative pathway species (Table S7). Additionally, eleven (out of 22) of these have *GAL* genes that are highly divergent in their sequences, indicating that they may have homologs that do not reach the sequence similarity threshold (Table S7). These yeasts could have very divergent, but still functional, *GAL* genes; their *GAL* genes may have been misannotated; or they have incomplete genomes that are missing the full sequences of the *GAL* genes. These yeasts may also require cryptic inducing conditions to test positive for growth on galactitol since they are often closely related to our documented alternative pathway species (Table S7).

The broader take-home message of our study is that machine learning approaches harbor great promise for studying the macroevolution of the genotype-phenotype map. The random forest algorithm used to analyze this dataset was very efficient in finding relevant genes and traits that predict growth on several carbon substrates with high accuracy, without requiring extensive manual parameter tuning. Part of its success is likely because we used the one-of-a-kind matrix of genomic, metabolic, and ecological data of the Y1000+ Project (6). While similar data matrices for other fungal or eukaryotic lineages are currently lacking, it would be fascinating to apply this type of analysis in clades with different morphologies, ecologies, or lifestyles than those of Saccharomycotina. While generation of data matrices equivalent to the one currently available for Saccharomycotina will undoubtedly require extensive effort and coordination, the potential for discovery is likely to be greater in lesser studied lineages.

Of course, how successful machine learning or any other genotype-phenotype association approach (52) will be for bridging genomic and phenotypic variation across macroevolutionary timescales will depend on numerous factors, including: the genetic architecture of the trait (oligogenic vs. polygenic); the degree to which the evolution of the trait is correlated with the evolution of other traits (univariate vs. multivariate); how often the trait has evolved (once vs. repeatedly); and whether the evolutionary mechanisms that contribute the trait are conserved (conserved vs. divergent). Oligogenic, univariate, repeatedly evolved traits that arise by the same evolutionary mechanisms will be the easiest to study. In certain respects, the *GAL* pathway fits these descriptions quite well; the ability to grow on galactose is encoded by a few genes (13), growth on galactose is only weakly correlated with growth on other traits (4), and the trait has been repeatedly gained and lost (15, 19). We therefore find it striking that machine learning enabled us to discover novel biology, namely the existence of an alternative pathway not previously known to be present in Saccharomycotina, in such a well-studied trait. When coupled with rich data, such as the treasure-trove of genomic, metabolic, and ecological data of the Y1000+ Project (6), we believe that machine learning approaches hold tremendous power to elucidate how genomic variation transforms into phenotypic variation across the tree of life.

Materials and Methods

Genomic data matrix

Using the KEGG (53, 54) and InterProScan (55) gene functional annotations generated by the Y1000+ Project (6), a data matrix was built with presence and absence of each unique KEGG Orthology (KO) and counts of each unique InterPro ID number in each genome. Each genome was its own row, and each unique KO (*N* = 5,043) or InterPro ID (*N* = 12,242) present in one or more of the 1,154 yeast genomes was its own column. A python script recorded the presence and absence of KO annotations (Table S1), the number of each InterPro ID for each genome (Table S2), and put them in the appropriate cells of the data matrix. Upon observing that accuracy was typically similar for predicting growth on 29 carbon sources between a random forest algorithm trained just on the KO dataset and the combined KO and InterPro dataset, the KO genomic dataset was used for all subsequent analyses, and the InterPro data was dropped from the genomic analyses following Figure 2. Comparison of our own *GAL* gene searches with the KO dataset revealed that *GAL1* was misannotated, and that the mutarotase and epimerase domains of *GAL10* were annotated separately by KEGG.

Metabolic data matrix

Our metabolic data matrix contained 122 traits from 893 yeast strains from 885 species in the subphylum. The list of traits included growth on different carbon and nitrogen sources, such as galactose, raffinose, and urea, as well as on environmental conditions, such as growth at different temperatures and salt concentrations (Table S3). The metabolic data were sourced from information available for each of the sequenced strains from the CBS strain database. These data were gathered from strains studied as part of the in the published descriptions of species, additional data on strains obtained by previous studies done in the Westerdijk Fungal Biodiversity Institute (CBS), or additional data provided by the depositors of the strains in the CBS culture collection. The data matrix contained metabolic data for 893/1,154 species. The percentage of missing data in the data matrix was 37.5% (40,906 missing values out of 108,946 total). Less thoroughly studied traits tended to have more missing data than more commonly found and/or thoroughly studied traits. For example, our data matrix included data on melibiose fermentation, which was estimated to be present in 12% (28/234) of yeasts, but only 26.2% (234/893 of strains have been tested for growth on this substrate. In contrast, our data matrix included data on galactose assimilation, which was estimated to be present in 64.2% (558/868), but 97.2% (868/893) of strains have been tested. Since there were 25 strains for which growth on galactose was not characterized, the total number of strains for which we have both genomic data and galactose assimilation data was 868.

Environmental data matrix and ontology

The isolation environments for 1,088 (94%) out of the 1,154 yeasts examined were gathered from strain databases, species descriptions, or from *The Yeasts: A Taxonomic Study* (6, 56). Strains without isolation environments either had been significantly domesticated via crossing or subculturing or were lacking information in our searches. Written descriptions of the environments were converted into a hierarchical trait matrix using a controlled vocabulary. The ontology was built with Web Protégé (https://webprotege.stanford.edu/), with six broader categories: animal, plant, environmental, fungal, industrial products, and victuals (food or drink). Within these categories, more specific controlled vocabulary annotations were connected to each strain: for example, an isolation environment reported as "*Drosophila hibisci* on *Hibiscus heterophyllus*" was associated in our ontology with the animal subclass "*Drosophila hibisci*" and the plant subclass "*Hibiscus heterophyllus*". This ontology was converted to a binary trait matrix containing all the unique environmental descriptors (Table S4). The same ontology was used in the recent Y1000+ manuscript (6), but that manuscript only considered the first subclass in subsequent analyses; our analyses here used all connections in the ontology for training a random forest algorithm.

Predicting growth on different carbon sources using machine learning algorithms trained on genomic, metabolic, and/or environmental data

To test whether we could predict growth on 29 different carbon sources from genomic, environmental, and/or (the rest of the) metabolic data, we used a random forest algorithm. These 29 traits were selected because they were measured in at least 743 strains and were present in 20%-80% of strains included in this analysis. For each trait, a random forest algorithm was trained separately on environmental, metabolic, or genomic datasets to evaluate the accuracy of prediction and identify the most important predictive features (Table S5).

We trained a machine learning algorithm built by an XGBoost (1.7.3) (57) random forest classifier (XGBRFClassifier()) with the parameters " max_depth=12 and n_estimators=100; all other parameters were in their default settings. The max_depth parameter specifies the depth of each decision tree, determining how complex the random forest will be to prevent overfitting while maintaining accuracy. The n_estimators parameter specifies the number of decision trees in the forest—after testing the increase in accuracy while increasing each of these parameters, we found that having a higher max_depth or more decision trees per random forest did not further increase accuracy.

The random forest algorithm was trained on 90% of the data, and used the remaining 10% for cross-validation, using the RepeatedStratifiedKFold and cross_val_score functions from the sklearn.model_selection (58) (1.2.1) package. Cross validation is a method for assessing accuracy involving 10 trials, each of which holds back a random 10% of the training data for testing (57, 58). The mean accuracy of the algorithm from this test was used for our in-depth xylose, sucrose, and galactose analyses, as those datasets were relatively balanced; that is, there were relatively similar numbers of strains that grew in these substrates (growers) and strains that did not grow in them (non-growers). For the analyses involving all 29 carbon substrates, we used balanced accuracy, which takes the mean of the true positive rate and the true negative rate, since there were unequal numbers of growers and non-growers in many of these substrates. For both measures, an accuracy value of 50% would be equivalent to randomly guessing.

Receiver Operator Characteristic (ROC) curves, which plot the true positive rate against the false positive rate, were also generated for each prediction analysis to visualize the accuracy of the algorithm in predicting growth on a given substrate—values of area under the curve (AUC) greater than 0.5 in these plots indicate better than random accuracy. We also used the cross_val_predict() function from Sci-Kit Learn separately to generate the confusion matrices; these matrices show the numbers of strains correctly predicted to grow or not grow on a specific carbon source (True Positives and True Negatives, respectively) and incorrectly predicted (False Positives, predicted to grow but do not; and False Negatives, not predicted to grow but do). This function also employs a 10-fold cross validation step, but it keeps track of which species are classified as True/False Positives and True/False Negatives during each of these 10 trials, while entering the final results into a confusion matrix. Top features were automatically generated by the XGBRFClassifier function using Gini importance, which uses node impurity (the amount of variance in growth on a given carbon source for strains that either have or do not have this trait/feature).

In each prediction analysis, we excluded from each training dataset growth and fermentation data for each of the 29 carbon sources under investigation. For example, we excluded growth on galactose and galactose fermentation from the training dataset for predicting growth on galactose; thus, the final metabolic data matrix used in the training contained data from 120 sources and conditions, instead of the total 122. Similarly, we excluded growth on sucrose and sucrose fermentation from the training dataset for predicting growth on sucrose; we excluded xylose and xylose fermentation from the training dataset for predicting growth on xylose. The

code used for these analyses is available at https://github.com/mcharrison95/RF_for_ML_GAL_paper.

GAL1, _GAL7_, _GAL10_, and _GAL102_ gene searches

To determine presence/absence of genes in the *GAL* pathway in each of the genomes of the 1,154 strains included in our study, we conducted sequence similarity searches for the *GAL1*, *GAL7*, *GAL102*, and *GAL10* genes using the jackhmmer function from the HMMER software, version 3.3.2 (59). Using the representative *GAL* gene sequences from the *Candida albicans* genome, jackhmmer searched for all hits above a similarity score of 200, which captured genes from all 12 Saccharomycotina taxonomic orders, and then used these results to build a new profile to search for the gene throughout the phylogeny. jackhmmer repeated this method until the results converged, which was three rounds for all genes except *GAL10*, which required five rounds, likely because the mutarotase and epimerase domains are part of the same protein in some yeast orders (e.g., Saccharomycetales and Serinales) but belong to two separate proteins (encoded by *GALM* and *GALE*, respectively) in others (e.g., Lipomycetales) (15, 19). In analyses where only the *GAL* gene dataset was used as genomic data, both the presence/absence and similarity score produced by jackhammer for *GAL1*, *GAL7*, and *GAL10* were included in the dataset; hits with similarity scores below 200 were considered absent and were entered as 0 (Table S6). As noted above, comparison of our own *GAL* gene searches with the KO dataset revealed that *GAL1* was misannotated, and that the mutarotase and epimerase domains of *GAL10* were annotated separately by KEGG.

Quantification of galactose utilization in strains lacking the _GAL_ pathway

To validate galactose utilization by certain strains lacking the *GAL* genes that were identified in our qualitative metabolic data matrix, we quantified growth and galactose consumption in liquid culture. Standard undefined yeast lab media was prepared as previously described (60). YPD medium for culturing yeasts contained 10 g/L yeast extract, 20 g/L peptone, 20 g/L glucose, and 18 g/L agar (US Biological). Cells were streaked onto YPD plates, and single colonies were picked. Cells were inoculated into 5 mL of YP (10 g/L yeast extract, 20 g/L peptone) + 2% galactose (Amresco) and grown to mid-log phase (48 – 55 hours depending on the strain, see Table S10 for further information) on a tissue culture wheel at room temperature. The optical density of the cells was measured at 600 nm (OD_{600}) using an OD600 DiluPhotometer (Implen). Cells were inoculated into 50 mL YP + 2% galactose at a starting OD_{600} 0.05 for all species except for the negative control species, *Saccharomycopsis malanga*, which was inoculated at starting OD_{600} 0.01 due to the low cell density caused by the absence of its *GAL* pathway. The cultures were shaken in non-baffled 150-mL Erlenmeyer flasks (Fisher Scientific) at 250 rpm at room temperature for seven days. 1 mL of culture was collected every 24 hours and spun down; 600 µL of supernatant were used for extracellular sugar quantification via high performance liquid chromatography and refractive index detection (HPLC-RID). OD_{600} readings were also taken at each 24-hour timepoint. All samples taken for HPLC-RID were stored at -20 ºC until the end of the experiment. Extracellular galactose concentrations were determined by HPLC-RID as previously described using a galactose standard (61, 62). The strain *S. cerevisiae gre3Δ::loxP-kanMX-loxP* (63) served as a positive control for galactose utilization because it has an intact *GAL* pathway; the deletion of *GRE3*, which encodes a promiscuous aldose reductase that could conceivably have some activity on galactose (64), also allowed this strain to serve as a negative control for the hypothesized oxidoreductive pathway. Galactose concentrations were expressed as g/L, and the results correspond to the mean value of biological triplicate timepoints. All extracellular galactose quantification data visualization was performed using R (v4.1.2) in the RStudio platform (v2022.07.01+554) and with the package ggplot2 (v3.4.2) (65, 66).

Assay for galactose- and NADPH-dependent enzymatic activity

To determine whether galactose utilization in strains lacking the *GAL* genes but able to grow in galactose occurred through a hypothesized oxidoreductive D-galactose pathway, we tested NADPH-dependent enzymatic activity on galactose as a sole carbon source. Yeast cells were

pregrown in YPD, single colonies were inoculated into 5 mL YP + 2% galactose, cultures were grown to mid-log phase, and they were inoculated into 50 mL YP + 2% galactose using the same methods as described above. *Candida duobushaemulonii*, *Candida ruelliae*, and *Ogataea methanolica* cells were harvested at mid-log phase along with their respective *S. cerevisiae gre3Δ::loxP-kanMX-loxP* negative controls for whole-cell lysate protein extraction using Y-PER (Thermo Fisher Scientific). 1 mL of culture was sampled, and cells were centrifuged at 3,000 x *g* at 4 ºC for 5 minutes. 250 mg of wet cell pellet were resuspended in 1,250 µL of Y-PER and homogenized by pipetting. The mixture was left to agitate at room temperature for 50 minutes to ensure successful cell lysis and soluble protein extraction. Cell debris was pelleted at 14,000 x *g* for 10 minutes at room temperature. Finally, 1 mL of supernatant was removed for analysis and protein concentration determination. Protein concentrations were determined using the Pierce BCA protein assay kit and protocol (Pierce Biotechnology), and absorbance at 562 nm was measured using The Infinite M1000 microplate reader (Tecan). Galactose-dependent enzymatic activity was determined by monitoring the oxidation of the cofactor NADPH to $NADP^+$ by absorbance measurement at 340 nm at 25 ºC (67). The assay mixture (200 µL) contained 200 mM Tris-HCl (pH 7.5), 5 mM of NADPH, 200 mM of galactose, 200 µg of undefined cell-free protein extract, and deionized water in 96-well plates (Corning 96 Well Clear Flat Bottom UV-Transparent). In addition, each assay contained a protein extract blank and a substrate (without galactose) blank to account for protein and substrate noise, cofactor degradation, and off-target cofactor oxidation. Enzyme assays were performed in biological quadruplicate. Data analyses and plots were performed and visualized using the methods described above.

Data availability statement

The supplementary dataset is available at https://doi.org/10.6084/m9.figshare.24855294. The code used to run the random forest algorithm is available at https://github.com/mcharrison95/RF_for_ML_GAL_paper. All Y1000+ Project genome sequence assemblies and raw sequencing data have been deposited in GenBank (6) and are available at the Figshare+ repository at https://doi.org/10.25452/figshare.plus.c.6714042.

Acknowledgments

We thank Tony Capra, the Hittinger Lab, the Rokas Lab, and Y1000+ Project team members for helpful discussions throughout the duration of this project; Trey K. Sato for the control strain of *S. cerevisiae*; and Mick McGee, Steve Karlen, and the GLBRC Metabolomics Facility for metabolite quantification. This work was performed using resources contained within the Advanced Computing Center for research and Education at Vanderbilt University in Nashville, TN. X.X.S. was supported by the National Science Foundation for Distinguished Young Scholars of Zhejiang Province (LR23C140001), the Fundamental Research Funds for the Central Universities (226-2023-00021), and the key research project of Zhejiang Lab (2021PE0AC04). This work was supported by the National Science Foundation (grants DEB-2110403 to C.T.H. and DEB-2110404 to A.R.). Research in the Hittinger Lab is also supported by the USDA National Institute of Food and Agriculture (Hatch Projects 1020204 and 7005101), in part by the DOE Great Lakes Bioenergy Research Center (DOE BER Office of Science DE–SC0018409, and an H. I. Romnes Faculty Fellowship (Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation). Research in the Rokas lab is also supported by the National Institutes of Health/National Institute of Allergy and Infectious Diseases (R01 AI153356), and the Burroughs Wellcome Fund.

References

1. C. T. Hittinger, J. L. Steele, D. S. Ryder, Diverse yeasts for diverse fermented beverages and foods. *Curr. Opin. Biotechnol.* **49**, 199–206 (2018).
2. A. Yaguchi, D. Rives, M. Blenner, New kids on the block: emerging oleaginous yeast of biotechnological importance. *AIMS Microbiol.* **3**, 227–247 (2017).
3. N. T. Case, *et al.*, The future of fungi: threats and opportunities. *G3 GenesGenomesGenetics* **12**, jkac224 (2022).
4. D. A. Opulente, *et al.*, Factors driving metabolic diversity in the budding yeast subphylum. *BMC Biol.* **16**, 26 (2018).
5. X.-X. Shen, *et al.*, Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum. *Cell* **175**, 1533-1545.e20 (2018).
6. D. A. Opulente, *et al.*, Genomic and ecological factors shaping specialism and generalism across an entire subphylum. 2023.06.19.545611 (2023).
7. C. T. Hittinger, *et al.*, Genomics and the making of yeast biodiversity. *Curr. Opin. Genet. Dev.* **35**, 100–109 (2015).
8. R. Riley, *et al.*, Comparative genomics of biotechnologically important yeasts. *Proc. Natl. Acad. Sci.* **113**, 9882–9887 (2016).
9. S. Ostergaard, L. Olsson, J. Nielsen, Metabolic Engineering of Saccharomyces cerevisiae. *Microbiol. Mol. Biol. Rev.* **64**, 34–50 (2000).
10. C. A. Brown, A. W. Murray, K. J. Verstrepen, Rapid Expansion and Functional Divergence of Subtelomeric Gene Families in Yeasts. *Curr. Biol.* **20**, 895–903 (2010).
11. M. Ptashne, A. Gann, *Genes and Signals*, 1st edition (Cold Spring Harbor Laboratory Press, 2001).
12. M. Johnston, A model fungal gene regulatory mechanism: the GAL genes of Saccharomyces cerevisiae. *Microbiol. Mol. Biol. Rev.* **51**, 458–476 (1987).
13. M.-C. Harrison, A. L. LaBella, C. T. Hittinger, A. Rokas, The evolution of the GALactose utilization pathway in budding yeasts. *Trends Genet.* **38**, 97–106 (2022).
14. X. Sun, *et al.*, Recognition of galactose by a scaffold protein recruits a transcriptional activator for the GAL regulon induction in Candida albicans. *eLife* **12**, e84155 (2023).
15. M. A. B. Haase, *et al.*, Repeated horizontal gene transfer of GALactose metabolism genes violates Dollo's law of irreversible loss. *Genetics* **217** (2021).
16. A. Venkatesh, A. L. Murray, A. Y. Coughlan, K. H. Wolfe, Giant GAL gene clusters for the melibiose-galactose pathway in Torulaspora. *Yeast* **38**, 117–126 (2021).
17. J. Boocock, M. J. Sadhu, A. Durvasula, J. S. Bloom, L. Kruglyak, Ancient balancing selection maintains incompatible versions of the galactose pathway in yeast. *Science* **371**, 415–419 (2021).
18. C. T. Hittinger, *et al.*, Remarkably ancient balanced polymorphisms in a multi-locus gene network. *Nature* **464**, 54–58 (2010).
19. J. C. Slot, A. Rokas, Multiple GAL pathway gene clusters evolved independently and by different mechanisms in fungi. *Proc. Natl. Acad. Sci.* **107**, 10136–10141 (2010).
20. C. A. Sellick, R. N. Campbell, R. J. Reece, "Chapter 3 Galactose Metabolism in Yeast—Structure and Regulation of the Leloir Pathway Enzymes and the Genes Encoding Them" in *International Review of Cell and Molecular Biology*, (Academic Press, 2008), pp. 111–150.
21. V. Brown, J. Sabina, M. Johnston, Specialized Sugar Sensing in Diverse Fungi. *Curr. Biol.* **19**, 436–441 (2009).
22. K. C. Gross, P. B. Acosta, Fruits and vegetables are a source of galactose: Implications in planning the diets of patients with Galactosaemia. *J. Inherit. Metab. Dis.* **14**, 253–258 (1991).
23. P. B. Acosta, K. C. Gross, Hidden sources of galactose in the environment. *Eur. J. Pediatr.* **154**, S87-92 (1995).
24. V. Marsilio, C. Campestre, B. Lanza, M. De Angelis, Sugar and polyol compositions of some European olive fruit varieties (Olea europaea L.) suitable for table olive purposes. *Food Chem.* **72**, 485–490 (2001).

25. A. Pontes, *et al.*, Tracking alternative versions of the galactose gene network in the genus Saccharomyces and their expansion after domestication. *iScience*, 108987 (2024).
26. C. K. Dalal, *et al.*, Transcriptional rewiring over evolutionary timescales changes quantitative and qualitative properties of gene expression. *eLife* **5**, e18981 (2016).
27. C. Ricci-Tam, *et al.*, Decoupling transcription factor expression and activity enables dimmer switch gene regulation. *Science* **372**, 292–295 (2021).
28. J. Zou, *et al.*, A primer on deep learning in genomics. *Nat. Genet.* **51**, 12–18 (2019).
29. B. M. Moore, *et al.*, Robust predictions of specialized metabolism genes through machine learning. *Proc. Natl. Acad. Sci.* **116**, 2344–2353 (2019).
30. A. S. Walker, J. Clardy, A Machine Learning Bioinformatics Method to Predict Biological Activity from Biosynthetic Gene Clusters. *J. Chem. Inf. Model.* **61**, 2560–2571 (2021).
31. O. Riedling, A. S. Walker, A. Rokas, Predicting fungal secondary metabolite activity from biosynthetic gene cluster data using machine learning. *Microbiol. Spectr.*, 2023.09.12.557468 (2023).
32. J. Zrimec, *et al.*, Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. *Nat. Commun.* **11**, 6141 (2020).
33. J. A. Capra, R. A. Laskowski, J. M. Thornton, M. Singh, T. A. Funkhouser, Predicting Protein Ligand Binding Sites by Combining Evolutionary Sequence Conservation and 3D Structure. *PLOS Comput. Biol.* **5**, e1000585 (2009).
34. W. Ma, *et al.*, A deep convolutional neural network approach for predicting phenotypes from genotypes. *Planta* **248**, 1307–1318 (2018).
35. S. Haridas, *et al.*, 101 Dothideomycetes genomes: A test case for predicting lifestyles and emergence of pathogens. *Stud. Mycol.* **96**, 141–153 (2020).
36. G. Biau, E. Scornet, A random forest guided tour. *TEST* **25**, 197–227 (2016).
37. X. Chen, H. Ishwaran, Random Forests for Genomic Data Analysis. *Genomics* **99**, 323–329 (2012).
38. J. R. Greenberg, N. P. Price, R. P. Oliver, F. Sherman, E. Rustchenko, Candida albicans SOU1 encodes a sorbose reductase required for L-sorbose utilization. *Yeast Chichester Engl.* **22**, 957–969 (2005).
39. D. J. Wohlbach, *et al.*, Comparative genomics of xylose-fermenting fungi for enhanced biofuel production. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 13212–13217 (2011).
40. R. L. Nalabothu, *et al.*, Codon Optimization Improves the Prediction of Xylose Metabolism from Gene Content in Budding Yeasts. *Mol. Biol. Evol.* **40**, msad111 (2023).
41. J. Meng, *et al.*, GalR, GalX and AraR co-regulate d-galactose and l-arabinose utilization in Aspergillus nidulans. *Microb. Biotechnol.* **15**, 1839–1851 (2022).
42. E. Fekete, *et al.*, The alternative D-galactose degrading pathway of Aspergillus nidulans proceeds via L-sorbose. *Arch. Microbiol.* **181**, 35–44 (2004).
43. R. Tanimura, A. Hamada, K. Ikehara, R. Iwamoto, Enzymatic synthesis of 2-keto-d-gluconate and 2-keto-d-galactonate from d-glucose and d-galactose with cell culture of Pseudomonas fluorescens and 2-keto-galactonate from d-galactono 1,4-lactone with partially purified 2-ketogalactonate reductase. *J. Mol. Catal. B Enzym.* **23**, 291–298 (2003).
44. L. Sun, *et al.*, Two-Stage Semi-Continuous 2-Keto-Gluconic Acid (2KGA) Production by Pseudomonas plecoglossicida JUIM01 From Rice Starch Hydrolyzate. *Front. Bioeng. Biotechnol.* **8**, 120 (2020).
45. T. Chroumpi, *et al.*, Detailed analysis of the D-galactose catabolic pathways in Aspergillus niger reveals complexity at both metabolic and regulatory level. *Fungal Genet. Biol.* **159**, 103670 (2022).
46. M. C. Kuang, *et al.*, Repeated Cis-Regulatory Tuning of a Metabolic Bottleneck Gene during Evolution. *Mol. Biol. Evol.* **35**, 1968–1981 (2018).
47. S. Suzuki, T. Matsuzawa, Y. Nukigi, K. Takegawa, N. Tanaka, Characterization of two different types of UDP-glucose/-galactose4-epimerase involved in galactosylation in fission yeast. *Microbiology* **156**, 708–718 (2010).

48. C. T. Hittinger, A. Rokas, S. B. Carroll, Parallel inactivation of multiple GAL pathway genes and ecological diversification in yeasts. *Proc. Natl. Acad. Sci.* **101**, 14144–14149 (2004).
49. M. Groenewald, *et al.*, A genome-informed higher rank classification of the biotechnologically important fungal subphylum Saccharomycotina. *Stud. Mycol.* (2023) https:/doi.org/10.3114/sim.2023.105.01 (July 11, 2023).
50. B. Seiboth, B. Metz, Fungal arabinan and l-arabinose metabolism. *Appl. Microbiol. Biotechnol.* **89**, 1665–1673 (2011).
51. T. Matsuzawa, *et al.*, New insights into galactose metabolism by Schizosaccharomyces pombe: Isolation and characterization of a galactose-assimilating mutant. *J. Biosci. Bioeng.* **111**, 158–166 (2011).
52. S. D. Smith, M. W. Pennell, C. W. Dunn, S. V. Edwards, Phylogenetics is the New Genetics (for Most of Biodiversity). *Trends Ecol. Evol.* **35**, 415–425 (2020).
53. M. Kanehisa, S. Goto, KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
54. M. Kanehisa, M. Furumichi, Y. Sato, M. Kawashima, M. Ishiguro-Watanabe, KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res.* **51**, D587–D592 (2023).
55. P. Jones, *et al.*, InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
56. C. Kurtzman, J. W. Fell, T. Boekhout, *The Yeasts: A Taxonomic Study* (Elsevier, 2011).
57. T. Chen, C. Guestrin, XGBoost: A Scalable Tree Boosting System in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16., (Association for Computing Machinery, 2016), pp. 785–794.
58. F. Pedregosa, *et al.*, Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
59. S. R. Eddy, A new generation of homology search tools based on probabilistic inference. *Genome Inform. Int. Conf. Genome Inform.* **23**, 205–211 (2009).
60. F. Sherman, Getting started with yeast. *Methods Enzymol.* **350**, 3–41 (2002).
61. M. S. Schwalbach, *et al.*, Complex Physiology and Compound Stress Responses during Fermentation of Alkali-Pretreated Corn Stover Hydrolysate by an Escherichia coli Ethanologen. *Appl. Environ. Microbiol.* **78**, 3442–3457 (2012).
62. S.-B. Lee, *et al.*, Crabtree/Warburg-like aerobic xylose fermentation by engineered Saccharomyces cerevisiae. *Metab. Eng.* **68**, 119–130 (2021).
63. L. S. Parreiras, *et al.*, Engineering and two-stage evolution of a lignocellulosic hydrolysate-tolerant Saccharomyces cerevisiae strain for anaerobic fermentation of xylose from AFEX pretreated corn stover. *PloS One* **9**, e107499 (2014).
64. N. Masuda, *et al.*, Neoadjuvant anastrozole versus tamoxifen in patients receiving goserelin for premenopausal breast cancer (STAGE): a double-blind, randomised phase 3 trial. *Lancet Oncol.* **13**, 345–352 (2012).
65. R: The R Project for Statistical Computing (July 20, 2023).
66. H. Wickham, *ggplot2: Elegant Graphics for Data Analysis* (Springer International Publishing, 2016).
67. R. M. Cadete, *et al.*, Exploring xylose metabolism in Spathaspora species: XYL1.2 from Spathaspora passalidarum as the key for efficient anaerobic xylose fermentation in metabolic engineered Saccharomyces cerevisiae. *Biotechnol. Biofuels* **9**, 167 (2016).

Figure legends

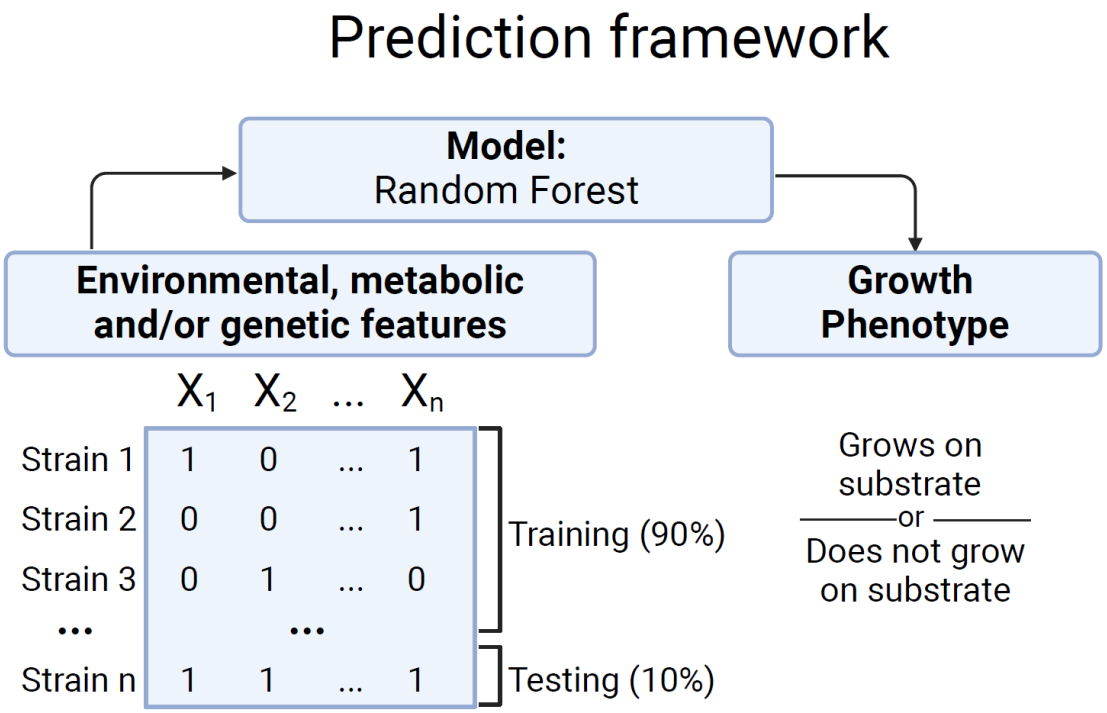


Figure 1. Workflow for machine learning prediction of how diet influences the evolution of primary metabolism in the subphylum Saccharomycotina. Using the phenotype of "grows on substrate" or "does not grow on substrate" for each yeast strain, we trained an XGBoost random forest algorithm on 90% of environmental, qualitative trait, and/or genetic features (893 strains containing 885 species). Using the 10% of remaining data, we tested model performance by looking at accuracy, confusion matrices, and ROC-AUC curves, and we repeated this assessment 9 more times using cross-validation. Feature importance was calculated using Gini importance as automatically generated by the XGBoost random forest algorithm. Created with BioRender.com.

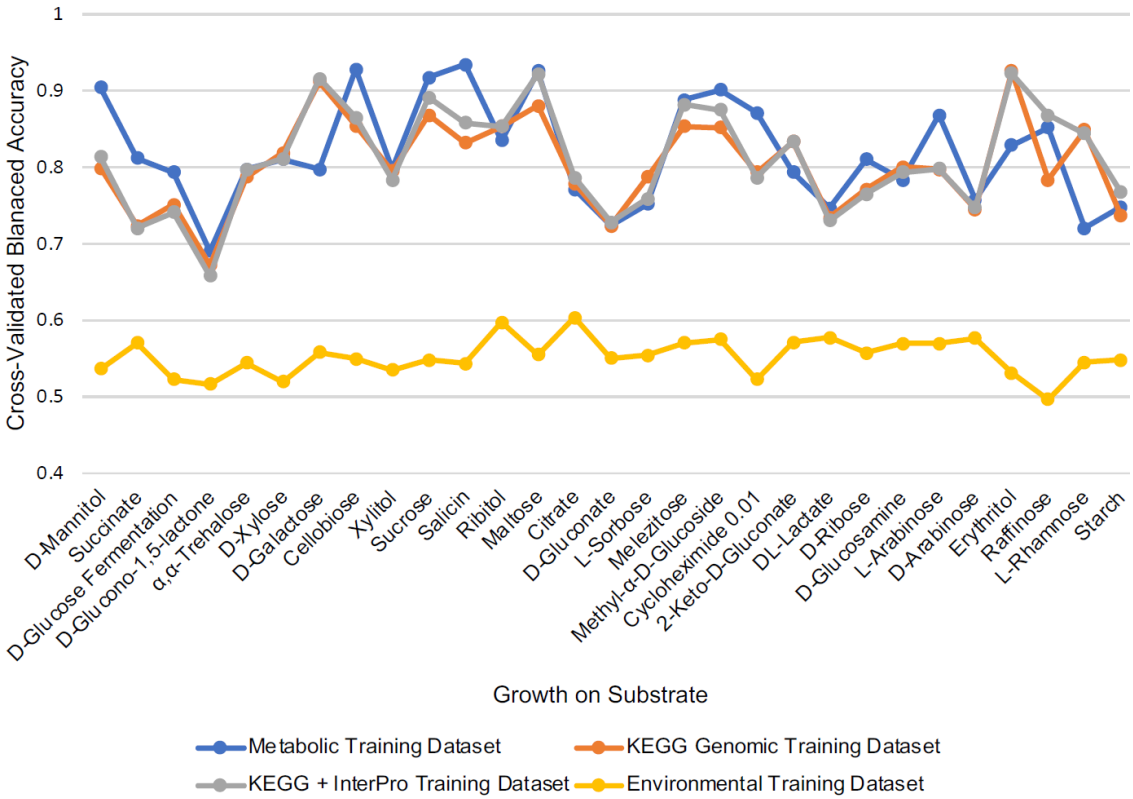


Figure 2. Prediction accuracy of growth on different substrates was high when the random forest algorithm was trained on metabolic data (blue) or genomic data (orange and grey) but low when the algorithm was trained on isolation environment data (yellow). Note that data on growth (and, where applicable, on fermentation) of the condition tested were removed prior to each analysis (e.g., prediction of growth on xylose from metabolic data was conducted using data for growth on all other substrates, but it excluded data for growth on xylose and xylose fermentation). Balanced accuracy was assessed by RepeatedStratifiedKFold (n_splits=10, n_repeats=3) after training the random forest algorithm on either the remainder of the metabolic data, the InterPro and/or KEGG genomic data matrices, or the environmental data. Traits are ordered from most frequent to least frequent in the dataset from left to right. The most important feature for each random forest algorithm, as well as the precision of the algorithm, is shown in the supplementary dataset (Supplementary Table 1).

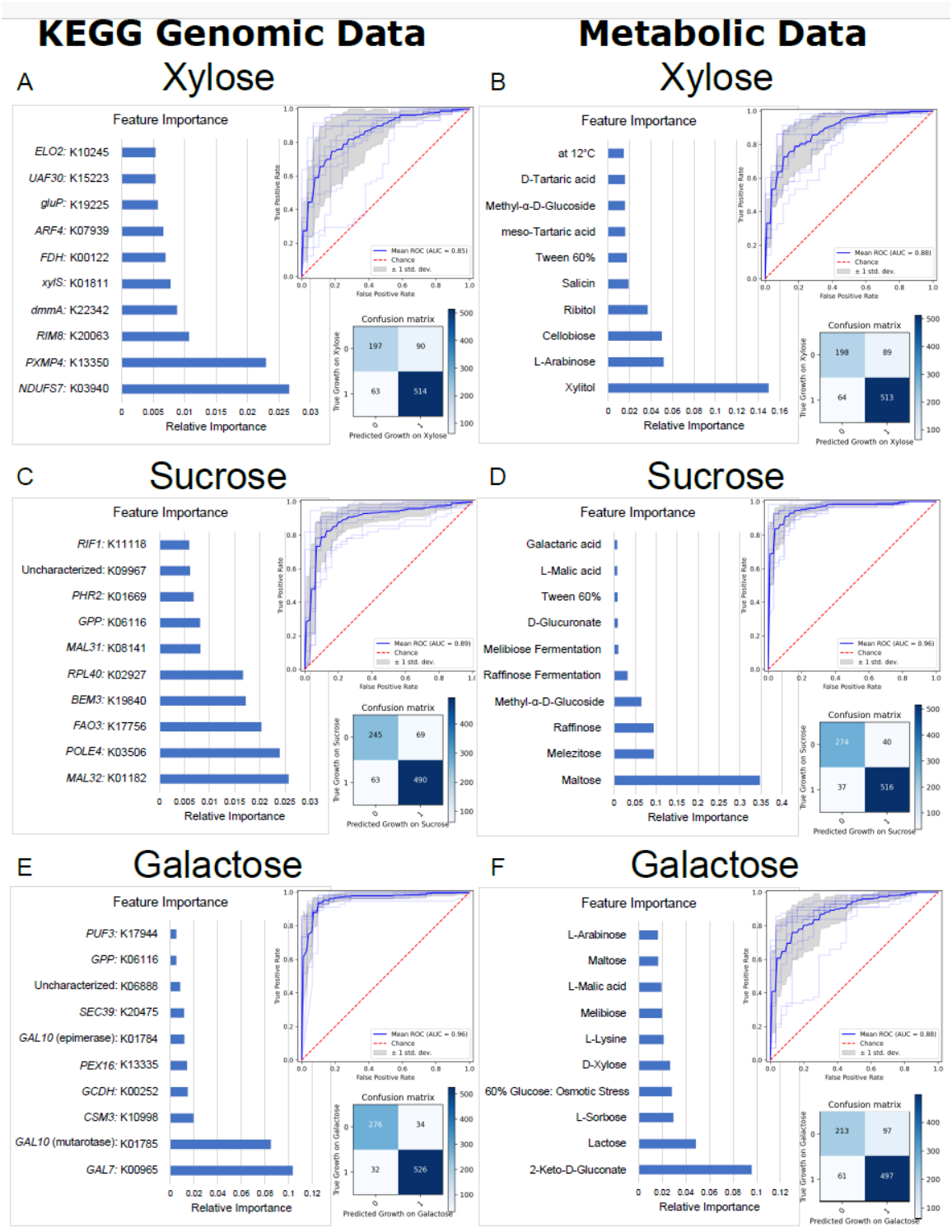


Figure 3. Prediction accuracy of growth on different sugars was high when the random forest algorithm was trained on genomic data (A, C, E), and similarly high when the algorithm was trained on metabolic data (B, D, F). Panels A and B: prediction of growth on xylose from genomic (A) or metabolic data (B). Panels C and D: prediction of growth on sucrose from genomic (C) or metabolic (D) data. Panels E and F: prediction of growth on galactose from

genomic (E) or metabolic (F) data. Note that data on growth (and, where applicable, on fermentation) of the carbon source tested were removed prior to each analysis (e.g., prediction of growth on xylose from metabolic data was conducted using data for growth on all other substrates and conditions, but it excluded data for growth on xylose and xylose fermentation). Also note that KEGG Orthology misannotated *GAL1,* likely leading *GAL1* to not be in the top features, and that the epimerase and mutarotase domains encoded by *GAL10* were annotated separately by this program. Accuracy is shown in the form of confusion matrices, which show strains predicted correctly to not grow on the sugar (true negatives, top left), strains predicted to grow on the sugar that do not (false positives, top right), strains correctly predicted to grow on the sugar (true positives, bottom right), and strains predicted to not grow on the sugar that do (false negatives, bottom left), as well as Receiver Operating Characteristic (ROC) curves, which show the true positive rate over false positive rate with changing classification thresholds. Feature importance graphs are also included to show the input features that are most useful for predicting growth on this sugar. XGBoost random forest was used to generate feature importance, and cross_val_predict() from sklearn.model_selection was used to generate confusion matrices. ROC curves were generated using the roc_curve function from sklearn.metrics. The prediction accuracies of growth on xylose, sucrose, and galactose from isolation environment data are shown in Supplemental Figure 1.

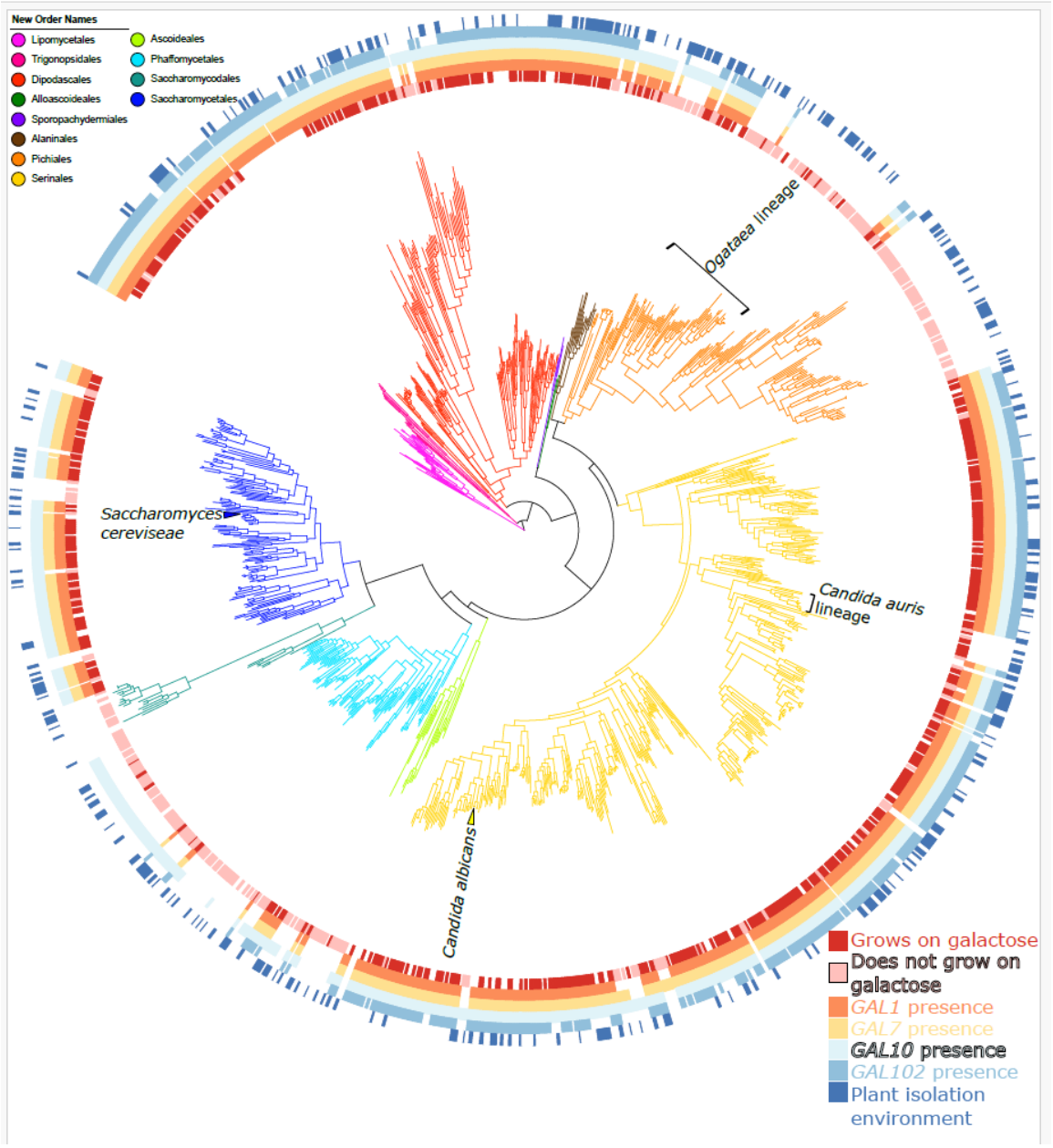


Figure 4. Distribution of *GAL* genes and plant isolation environments across the yeast phylogeny. The ability of the different strains to grow on galactose, the presence of genes *GAL1*, *GAL7*, *GAL10*, and *GAL102*, and whether they were isolated from plant environments are plotted as circles around the yeast phylogeny. Strain names are omitted for easier visualization, but they can be found in Figure S2. The colors of the different branches of the phylogeny correspond to the 12 taxonomic orders (49).

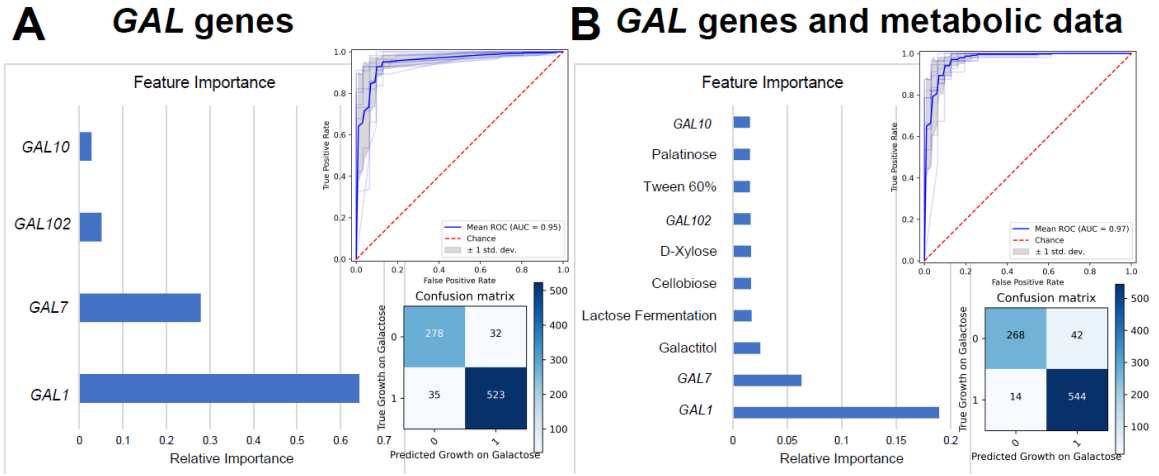
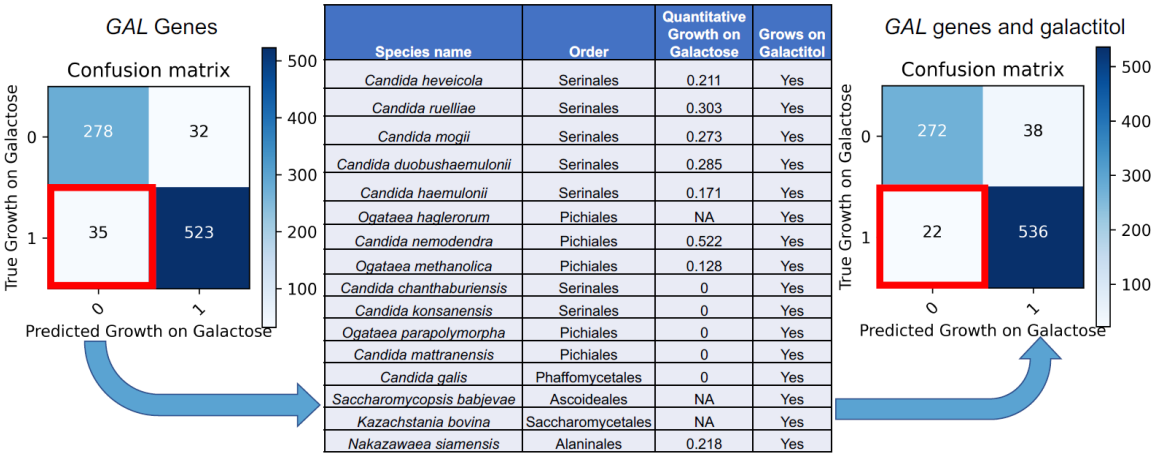


Figure 5. ***GAL* gene presence / absence and ability to grow on galactitol are highly predictive of growth on galactose across the subphylum Saccharomycotina. A.** Using the presence / absence patterns of the genes *GAL1, GAL7, GAL10,* and *GAL102* as input data, the XGBoost random forest algorithm predicted growth on galactose with high accuracy, as shown by the confusion matrix, the ROC/AUC curve, and the individual feature importance. **B.** Using both the presence / absence patterns of *GAL* genes (from panel A) and metabolic data, the algorithm predicted growth on galactose with even higher accuracy, shown by the confusion matrix, the ROC/AUC curve, and the individual feature importance. Note that, after *GAL1*, *GAL7*, and *GAL102* genes, growth on galactitol is the next most important feature for predicting growth on galactose.



Species name	Order	Quantitative Growth on Galactose	Grows on Galactitol
Candida heveicola	Serinales	0.211	Yes
Candida ruelliae	Serinales	0.303	Yes
Candida mogii	Serinales	0.273	Yes
Candida duobushaemulonii	Serinales	0.285	Yes
Candida haemulonii	Serinales	0.171	Yes
Ogataea haglerorum	Pichiales	NA	Yes
Candida nemodendra	Pichiales	0.522	Yes
Ogataea methanolica	Pichiales	0.128	Yes
Candida chanthaburiensis	Serinales	0	Yes
Candida konsanensis	Serinales	0	Yes
Ogataea parapolymorpha	Pichiales	0	Yes
Candida mattranensis	Pichiales	0	Yes
Candida galis	Phaffomycetales	0	Yes
Saccharomycopsis babjevae	Ascoideales	NA	Yes
Kazachstania bovina	Saccharomycetales	NA	Yes
Nakazawaea siamensis	Alaninales	0.218	Yes

Figure 6. Adding the galactitol growth data to presence / absence of the *GAL* genes increased prediction accuracy by correctly classifying several false negatives as true positives. On the left is the confusion matrix for predicting growth on galactose using just *GAL1*, *GAL7, GAL10*, and *GAL102* presence / absence. Note the presence of 35 false negatives; the algorithm predicted that these 35 species would be unable to grow on galactose because they lack the *GAL* genes, but they are known to grow on galactose. When the metabolic trait "Growth on Galactitol" was added to the training data, 16 of these species were then correctly predicted to grow on galactose and were moved to the "True Positive" category, while 19 remained false negatives. Three additional species that have low sequence similarity scores for the presence of *GAL* genes in its genome (*Metschnikowia kofuensis*, *Kuraishia piskuri*, and *Wickerhamomyces subpelliculosus*) also became new false negatives, bringing the total up to 22 false negatives and 536 true positives, as shown in the confusion matrix on the right. The taxonomy (order)

(Groenewald et al. 2023), quantitative growth on galactose (which is normalized to growth on glucose), and qualitative ability to grow on galactitol for these 15 species are listed in the table. Additionally, it is worth noting that one of the species (*Nakazawaea siamensis*) that was a false negative and became a true positive has *GAL* genes with low sequence homology—with the addition of galactitol data, on which it does grow, it was then correctly predicted to grow on galactose.

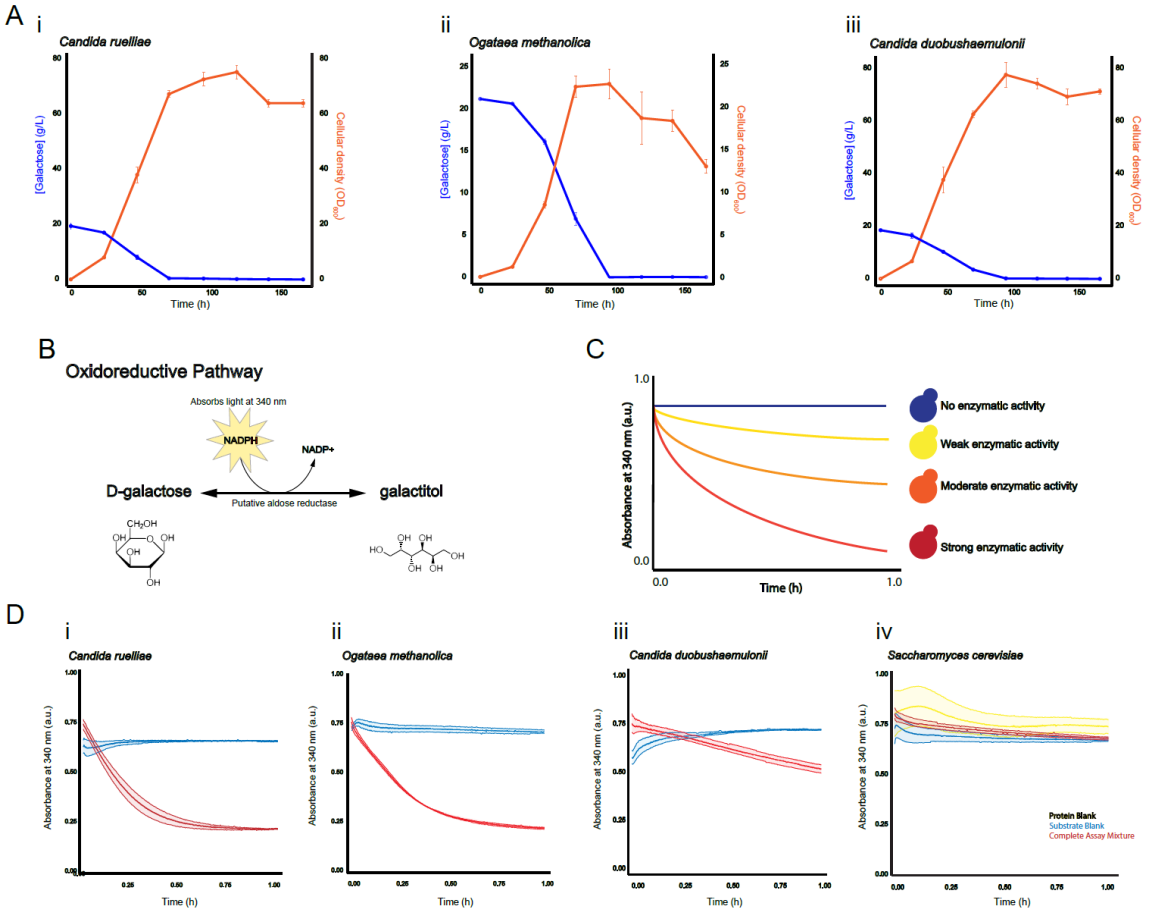


Figure 7. All three species showed galactose consumption and enzymatic activity on galactose. A. Average and standard deviation across three biological replicates of galactose concentrations present in medium with galactose as the sole carbon source (blue) and OD_{600} growth measurements (orange) for *C. ruelliae* (i), *O. methanolica* (ii), and *C. duobushaemulonii* (iii) over 168 hours. **B.** Schematic diagram of the first step of a hypothesized oxidoreductive galactose pathway using an aldose reductase to reduce galactose to galactitol by oxidizing NADPH to $NADP^+$. **C.** Illustration of the expected results for different levels of enzymatic activity. As the amount of NADPH present in the assay mixture decreases, absorbance at 340 nm decreases. **D.** Average and standard deviation across four biological replicates of NADPH absorbance at 340 nm over time comparing the complete assay mixture (red) to a substrate blank with no galactose added (blue) for *C. ruelliae* (i), *O. methanolica* (ii), *C. duobushaemulonii* (iii), and *S. cerevisiae* (iv). The same protein blanks (yellow) were used for all species included in the enzyme assay since each replicate of the enzyme assay included all four species on one 96-well plate, and the protein blank possessed reagents that were the same across all species (Tris-HCl, galactose, NADPH, and deionized water).