Exploring Saccharomycotina Yeast Ecology Through an Ecological Ontology Framework

Marie-Claire Harrison[1], Dana A. Opulente[2], John Wolters[3], Xing-Xing Shen[1], Xiaofan Zhou[1], Marizeth Goenewald[4], Chris Hittinger[3], Antonis Rokas[1], and Abigail LaBella[5]

[1]Vanderbilt University
[2]Villanova University
[3]University of Wisconsin-Madison
[4]Westerdijk Fungal Biodiversity Institute
[5]The University of North Carolina at Charlotte

July 16, 2024

Abstract

Yeasts in the subphylum Saccharomycotina are found across the globe in disparate ecosystems. A major aim of yeast research is to understand the diversity and evolution of ecological traits, such as carbon metabolic breadth, insect association, and cactophily. This includes studying aspects of ecological traits like genetic architecture or association with other phenotypic traits. Genomic resources in the Saccharomycotina have grown rapidly. Ecological data, however, are still limited for many species, especially those only known from species descriptions where usually only a limited number of strains are studied. Moreover, ecological information is recorded in natural language format limiting high throughput computational analysis. To address these limitations, we developed an ontological framework for the analysis of yeast ecology. A total of 1,088 yeast strains were added to the Ontology of Yeast Environments (OYE) and analyzed in a machine-learning framework to connect genotype to ecology. This framework is flexible and can be extended to additional isolates, species, or environmental sequencing data. Widespread adoption of OYE would greatly aid the study of macroecology in the Saccharomycotina subphylum.

TITLE: Exploring Saccharomycotina Yeast Ecology Through an Ecological Ontology Framework

RUNNING TITLE: A Yeast Ecological Ontology

AUTHORS: Marie-Claire Harrison[1], Dana A. Opulente[2], John F. Wolters[3], Xing-Xing Shen[4], Xiaofan Zhou[5], Marizeth Groenewald[6], Chris Todd Hittinger[7], Antonis Rokas[8], Abigail Leavitt LaBella[9^]

AUTHOR AFFILIATIONS:

1. Department of Biological Sciences, Vanderbilt University, Nashville, TN 37235, USA; Evolutionary Studies Initiative, Vanderbilt University, Nashville, TN 37235, USA
2. Biology Department Villanova University, Villanova, PA 19085, USA; Laboratory of Genetics, DOE Great Lakes Bioenergy Research Center, Wisconsin Energy Institute, Center for Genomic Science Innovation, J. F. Crow Institute for the Study of Evolution, University of Wisconsin-Madison, Madison, WI 53726, USA
3. Laboratory of Genetics, DOE Great Lakes Bioenergy Research Center, Wisconsin Energy Institute, Center for Genomic Science Innovation, J. F. Crow Institute for the Study of Evolution, University of Wisconsin-Madison, Madison, WI 53726, USA
4. 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
5. Guangdong Province Key Laboratory of Microbial Signals and Disease Control, Integrative Microbiology Research Center, South China Agricultural University, Guangzhou 510642, China
6. Westerdijk Fungal Biodiversity Institute, 3584 CT Utrecht, The Netherlands
7. Laboratory of Genetics, DOE Great Lakes Bioenergy Research Center, Wisconsin Energy Institute, Center for Genomic Science Innovation, J. F. Crow Institute for the Study of Evolution, University of Wisconsin-Madison, Madison, WI 53726, USA
8. Department of Biological Sciences, Vanderbilt University, Nashville, TN 37235, USA; Evolutionary Studies Initiative, Vanderbilt University, Nashville, TN 37235, USA
9. Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, North Carolina Research Campus, Kannapolis NC 28223, USA; Center for Computational Intelligence to Predict Health and Environmental Risks (CIPHER), University of North Carolina at Charlotte, 9201 University City Boulevard, Charlotte, NC, 28233, USA

^ Corresponding Author: alabell3@charlotte.edu

AUTHOR CONTRIBUTIONS:

MCH: Designed ontology

DAO, JFW, XXS, XZ, MG, CTH, AR provided computational support and reagents

ALL designed and implemented computational analyses, managed data, prepared figures, wrote the manuscript and supervised the project.

All authors provided comments on the manuscript.

KEYWORDS: controlled vocabulary, dynamic, formal, isolation environment, statistical enrichment, macroecology

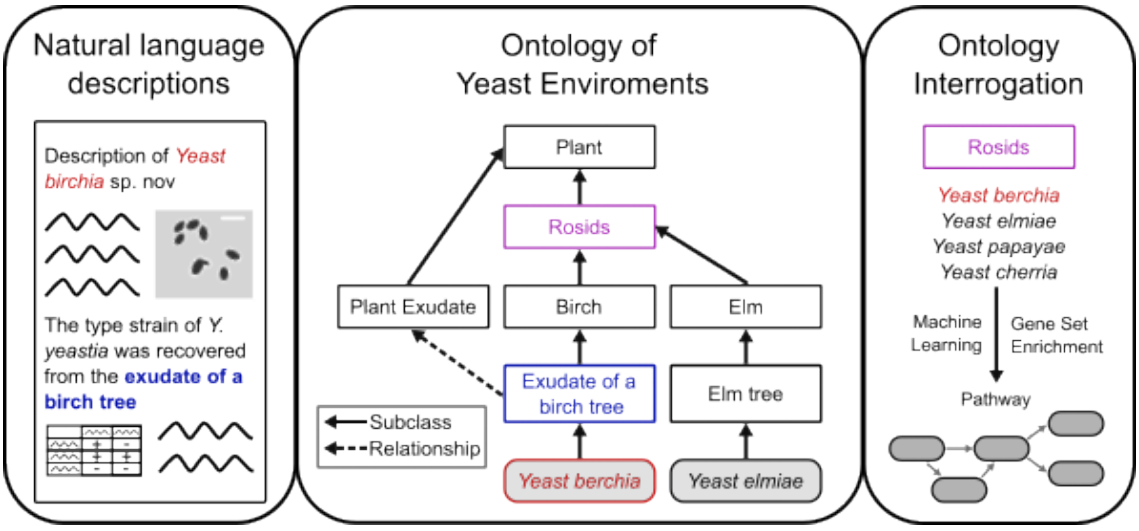
TAKE AWAYS

- Ontological frameworks allow high throughput analysis of ecological data
- We established a formal Ontology of Yeast Environments
- The Ontology of Yeast Environments describes isolation environments for 1,088 strains
- Coupled with genomic data, analysis of the ontology reveals gene-environment associations

ABSTRACT

Yeasts in the subphylum Saccharomycotina are found across the globe in disparate ecosystems. A major aim of yeast research is to understand the diversity and evolution of ecological traits, such as carbon metabolic breadth, insect association, and cactophily. This includes studying aspects of ecological traits like genetic architecture or association with other phenotypic traits. Genomic resources in the Saccharomycotina have grown rapidly. Ecological data, however, are still limited for many species, especially those only known from species descriptions where usually only a limited number of strains are studied. Moreover, ecological information is recorded in natural language format limiting high throughput computational analysis. To address these limitations, we developed an ontological framework for the analysis of yeast ecology. A total of 1,088 yeast strains were added to the Ontology of Yeast Environments (OYE) and analyzed in a machine-learning framework to connect genotype to ecology. This framework is flexible and can be extended to additional isolates, species, or environmental sequencing data. Widespread adoption of OYE would greatly aid the study of macroecology in the Saccharomycotina subphylum.

GRAPHICAL ABSTRACT



INTRODUCTION

The importance of yeast ecology

Over the past 400 million years, the yeasts in the subphylum Saccharomycotina (hereafter referred to as yeasts) spread across Earth, adapting to nearly every biome available (Kurtzman et al., 2011; Shen et al., 2020). The diversity of biotic and abiotic features in these global environments profoundly influenced the diversification and evolution of over 1,000 species of yeasts. It led to the evolution of varied genome content, metabolic capabilities, and phenotypic traits (Shen et al., 2018). Yeasts are now critical components of many different scientific realms: they are used in biotechnology as biofuel and heterologous protein producers (Riley et al., 2016); they play an essential role in the global food supply as plant pathogens, food, and beverage producers (Hittinger et al., 2018), and spoilage yeasts (Loureiro & Querol, 1999); and they impact human health as commensal (Suhr & Hallen-Adams, 2015) and pathogenic (Bidaud et al., 2018) components of the mycobiome.

The environments in which yeast thrive are as varied as the yeasts themselves. They are predicted to be most commonly found in mixed montane forests in temperate climates. However, yeasts have been sampled directly from the atmosphere, including from clouds (Vaïtilingom et al., 2012). In the aquatic realm, yeasts can be found in very high densities across freshwater, marine, and deep-sea environments (Nagahama, 2006). Within the deep-sea, yeasts have been found at deep-sea hydrothermal vents (Keeler et al., 2021), cold seeps (Nagano et al., 2014), and whale falls (Nagano et al., 2020). In the Arctic, yeasts have been isolated from seawater, subglacial ice, and brine puddles on sea ice (Butinar et al., 2011). On land, yeasts are found to be associated with abiotic substrates and living or dead organisms. Abiotic environments that host yeasts include soil (Botha, 2011), caves (Cunha et al., 2020), and rock surfaces (Selbmann et al., 2014).

Yeasts have also evolved intimate relationships with many different organisms. Yeasts, plants, and insects form complex systems where some or all the partners benefit. This includes the well-known cactus-yeast-*Drosophila* (Goncalves et al., 2023; Starmer & Fogleman, 1986) and flower-yeast-beetle systems (Blackwell, 2017). Other animals from which yeasts have been isolated include cows (Brejova et al., 2019), horses, chickens, bats, apes, and cats (Kurtzman et al., 2011). Yeasts play a major role in the digestive tracts of animals ranging from insects (Stefanini, 2018) to humans (Perez, 2021). In association with plants, yeasts are found on leaves (Slavikova et al., 2007), in plant exudates (Bowles & Lachance, 1983), and associated with roots (Sarabia et al., 2017). Yeasts also play a major role in the environment as decomposers of plant matter (Cadete et al., 2017). This list is not exhaustive, but it demonstrates the breadth of niches that yeasts inhabit.

Yeasts from these varied habitats exhibit different, likely adaptive traits. Yeasts isolated from cold seeps in the deep sea are adapted to low temperatures (Nagano et al., 2014). Yeasts isolated from mammalian digestive tracts can resist stressors, such as the immune system (Rosenbach et al., 2010). A better understanding of where yeasts reside and their ecological

niche breadths will allow us to test hypotheses regarding how their diverse ecological traits evolved, what yeast traits might emerge in the future, and what intrinsic or extrinsic factors have shaped their observed patterns of diversity in species across the yeast subphylum.

Uncovering genetic variants associated with ecological traits of yeast species remains a major challenge. Traditionally, researchers identify a trait and subsequently identify the genetic features that influence it. For example, the beak morphology of Darwin's finches is associated with variation in bone morphogenic protein 4 (BMP4) (Abzhanov et al., 2004). Identifying genetic contributors to ecological traits in microbes can be challenging due to sampling limitations, unknown genetic backgrounds, and complex phenotype-environment interactions (Brettner et al., 2022), even for well-characterized traits. For example, the ability of yeasts to produce and accumulate ethanol under aerobic conditions (the Crabtree/Warburg Effect) is associated with multiple genetic changes (Postma et al., 1989) and arose approximately 125-150 million years ago (Hagman & Piskur, 2015). Did microbial competition lead to this innovation? If so, under what specific conditions or environment did this trait arise? Previous analyses cannot confidently identify the forces shaping this trait due to the evolutionary time scale and lack of information about the ecological niche of extant yeasts (Hagman & Piskur, 2015). The known ecological data for Crabtree/Warburg-positive Saccharomycetaceae are highly varied. *Tetrapisispora phaffii* has been isolated once from African soil in the 1960s (Kurtzman et al., 2011). Conversely, *Kluyveromyces marxiaunus* has been isolated from foods, beverages, decaying plant tissue, and insects (Kurtzman et al., 2011). Given this data, we cannot make any clear connections between ecology and the Crabtree/Warburg Effect, let alone its adaptive significance. In other cases, different yeast species may share a trait, but the underlying genetic associations are not the same. For example, while most yeasts utilize the Leloir pathway to metabolize D-galactose, some yeasts appear to utilize an alternative oxidoreductive D-galactose pathway (Harrison et al., 2024). Conversely, many yeasts contain the enzymes necessary to metabolize xylose but are unable to grow on xylose in a laboratory setting (Nalabothu et al., 2023). These features—long evolutionary time scales, limited ecological data, complex genetic traits, and more—make traditional ecological studies difficult.

One approach that addresses some of the issues noted above is "Reverse Ecology," in which traits and their underlying genetic variation are inferred directly from genomic information (Levy & Borenstein, 2012). There are vast genomic resources available in yeasts, from thousands of strains within a species (Peter et al., 2018) to a genome for nearly every known yeast species (Opulente et al., 2024). This latter species-level dataset, known as the Y1000+ Project (http://y1000plus.org) dataset, provides genomes for 1,154 yeast strains from 1,051 species and, importantly for reverse ecology, phenotypic and ecological data. Yeast researchers have already begun to interrogate diverse ecological traits and link ecology or habitat with specific yeast traits and underlying genome variation (Cavalieri et al., 2022). Yeasts associated with fruits, fermented substrates, and juices are more likely to have the genomic capability to ferment both glucose and sucrose (Opulente et al., 2018). Cacti-associated yeasts exhibit elevated thermotolerance levels associated with increased evolution rates in cell envelope genes (Goncalves et al., 2023). Yeasts associated with dairy environments have genomic changes related to an increased growth rate on galactose media (LaBella et al., 2021). The dataset size

allows the utilization of big-data methods, such as machine learning and phylogenomic approaches. However, our current ecological data limit the application of the vast genomic and phenotypic data to address pressing ecological questions.

The ecology of yeasts and other microbes can be understood either through direct observation of the organisms in their natural environments or through inference of their potential habitats based on known traits and general ecological principles (Starmer & Lachance, 2011). We will focus here on the inference of yeast ecology from their isolation environments. Large-scale databases, such as the Global Biodiversity Information Facility (*GBIF: The Global Biodiversity Information Facility (2024)*) and GlobalFungi (Vetrovsky et al., 2020), provide such data, but they do so for a relatively small number of species. For example, a recent study identified records for 186 yeast species, which amounts to only ~15% of the described species (David et al., 2024). Metagenomic studies are beginning to enable the identification of yeasts from environmental DNA sampling. For example, a study identified the diversity of seven *Saccharomyces* species across elevations and tree habitats (Alsammar et al., 2019). Similarly, a metagenomic study of human cancer samples revealed evidence of 67 Saccharomycotina yeasts but could only identify the species of 23 of these (Narunsky-Haziza et al., 2022). The recent boom in yeast genome sequencing will further allow the identification of more yeasts in metagenomic studies. We anticipate these databases will continue to grow and capture more yeast ecology; capturing this information in digital formats that are consistent across studies will be key for large-scale studies of yeast ecology. In the meantime, there are bountiful opportunities to construct the computational framework for synthesis to leverage the currently available ecological information in novel ways that enable big data analysis.

Ecological Data & Bio-Ontologies

Ecological data are recorded during the collection of yeasts and documented in species descriptions. According to the current guidelines, species descriptions should include, “A clear statement of the geographic origin and habitat of all isolates” (Lachance, 2020). Ideally, this statement would include precise geographic information, detailed substrate description, temperature at the time of collection, and substrate pH. Recorded ecological data, especially historical data, rarely include all these features. In some cases, the data provided are sparse, such as ”rotting wood samples were collected in the Sanctuary of Caraça” (Morais et al., 2013). Other descriptions are highly detailed, such as ”larvae of *Anastrepha mucronata* (Diptera: Tephritidae) collected from ripe fruit of *Peritassa campestris* (‘Bacupari’, Hippocrateaceae)... in the Cerrado ecosystem of the state of Tocantins, Brazil” (Rosa et al., 2006). It is difficult to identify what information might be useful at the time of collection, especially without a universal language to describe environments. Even when detailed information is recorded, it must be re-recorded in a machine-readable format for high-throughput analyses.

Ontologies are an important framework used to transform information described in natural language into a format that allows integration across methods, technologies, and applications (Hastings, 2017). Natural language, simply the language used by humans to communicate, is rife with words with multiple meanings and other complexities that make biological

interpretations difficult. For example, the word “tree” does not refer to any specific monophyletic group of species—the word tree is used in reference to angiosperms, gymnosperms, and even palms. There is also no universally recognized age at which a sapling should be referred to as a tree or the height at which a shrub transitions to a tree. Therefore, it is reasonable to assume that the word tree represents many different ecological niches. Even species names do not always represent evolutionary relatedness. In Saccharomycotina yeasts, the generic name *Candida* has been used in four different orders, with 32% outside the lineage containing *Candida albicans* (Opulente et al., 2024). Ontologies, like phylogenies, allow us to define precise relationships between biological entities, which allows systematic data analysis and generates a dynamic but controlled vocabulary by which scientists can communicate.

Biological ontologies, also known as bio-ontologies, have become a key resource for scientists. The most popular bio-ontology is the Gene Ontology (GO) (Ashburner et al., 2000). The GO framework consists of three independent ontologies that use dynamic, controlled vocabularies to capture our current knowledge of the molecular functions, cellular components, and biological processes of genes. The success of the GO led to the development of the Open Biomedical Ontologies (OBO) (Smith et al., 2007), which provides best practices, tutorials, and tools for the development of ontologies ranging from Anatomy Ontology (Haendel et al., 2008) to the Zebrafish Phenotype Ontology (Van Slyke et al., 2014). In total, there are 600 ontologies currently listed in the OBO.

Another set of ontologies has been developed specifically to address evolutionary and ecological hypotheses. The Semantics for Comparative Analysis of Trait Evolution (SCATE) was developed to represent complex traits recorded in natural language format as ontologies for evolutionary analysis (Dahdul et al., 2017). This work builds on the success of Phenoscape (http://kb.phenoscape.org), which is an ontology-driven resource aimed at linking phenotypes across fields of biology. It has been used to identify candidate genes associated with phenotypes in fishes (Edmunds et al., 2016). There is also The Environment Ontology which describes environments ranging from ecosystems to planets and even astronomical bodies (Buttigieg et al., 2013). This ontology contains some terms that apply to yeasts, such as “wetland area,” but it cannot account for the many yeasts whose environment is another organism, such as the gut of a beetle. Therefore, the current biological, evolutionary, and ecological ontologies do not fully capture the breadth of yeast environments.

The extensive breadth of environments where yeasts are found necessitated a new ontology. There are bio-ontologies currently available for natural environments (Buttigieg et al., 2013), human anatomy (Haendel et al., 2008), food (Dooley et al., 2018), and plants (Jaiswal et al., 2005). Yeasts are found in all these environments and many more. Moreover, the ecology of some yeasts involves the close relationship between multiple environments. This includes the well-characterized cactus-yeast-*Drosophila* and flower-yeast-beetle systems (Starmer & Lachance, 2011). To address the specific challenges of studying yeast ecology, we constructed a new yeast environment ontology using the guiding principles outlined in the Ontology Development 101 (Noy & McGuinness, 2001) provided by the team that manages the ontology visualization tool Protégé (Musen & Protege, 2015). The ontology was constructed as part of the

Y1000+ Project and was used in the flagship publication of the 1,154 yeast genomes (Opulente et al., 2024). In this article, the ontology was used to identify overlapping isolation environments between metabolic specialist and generalist yeasts. We will refer to this ontology as the Ontology of Yeast Environments (OYE). Below, we will outline the steps for the construction of the ontology.

Construction of the Ontology of Yeast Environments (OYE)

Ontologies are comprised of classes, metadata, relations, and axioms stored in a common file format. We will use a beetle to illustrate these ideas. Classes are the most basic unit and are the hierarchical categories into which observations are placed. We could define “beetle” as a class. Metadata is any information stored within a class and could contain information like a written description. For example, we may include metadata, such as “Beetles are insects with defining features such as wing cases. There are likely millions of species of beetles.” Relations or modifiers connect classes to each other in the ontology and can include connections, such as “is a part of” to “has function”. We could connect the two classes, “beetle” and “wing-cases”, using a relationship called “is a part of.” Axioms are the rules that constrain classes. All members of the class “beetle” are also members of the class “insect” which makes that an axiom. We will refer to subclasses as any class connected by this type of axiom. This structure allows for flexibility and high-throughput computational analyses. These principles were used in the construction of the OYE

Step 1 was to identify key terms to guide the construction of the ontology. We collected the isolation environment in a natural language form from species descriptions or fungal collections for each strain in our set of 1,154 Saccharomycotina yeast strains. In total, we were able to identify information for 1,088 yeasts (Supplementary Data s6 from (Opulente et al., 2024)). The information was matched to the strain level to account for the possibility of within-species variation in associations between ecology and genome, such as polymorphisms found in the *GAL*actose metabolism pathway (Hittinger et al., 2010; Lee et al., 2017; Pontes et al., 2024).

In step 2, we reviewed the isolation environment information and created the most general exclusive classes for environments – animal, plant, environmental, fungal, industrial products, and victuals (food or drink). We also identified subclasses within these classes, such as *type*, *part*, and *product* (Figure 1.) A *type* is a specific instance of the category. For example, a hexapod is a *type* of arthropod, which is a *type* of animal. A *part* is a specific region of that category, such as the intestine, which is an internal *part* of the animal which is a *part* of an animal. Finally, a *product* is a material that originates from the type but can be collected or separated. For example, feces are a *product* of animals.

***Figure 1: Ontology subset describing the isolation environment of* Metschnikowia mauinuiana.** *Each box represents a distinct class in the ontology. Each class is a subclass of a single class higher-up in the ontology. There are two relational properties shown in the figure (green and yellow arrows) that describe relationships between classes. The strain of* M. mauinuiana *shown is an instance (red arrow) of the specific environment from which it was isolated.*

In step 3, we identified important features that may apply to some, but not all, of these environments, such as an association with microbes and the state of matter. These features have sub-categories, such as fermented as a sub-category of microbial association. We also outlined the modifier and relational properties that connect our categories. Many secondary associations exist between categories identified in the isolation environments, such as an insect found on a specific plant. Therefore, we created relational properties, such as “is from animal on plant.” We created modifier properties, such as “has microbe association”, to identify the relationship between our categories and the features. This step allowed us to define our ontology's scope and general structure.

Step 4 was to define the class hierarchy. We used the Web Protégé application to allow for collaborative work and visualization. The highest level of the hierarchy was split into exclusive classes: animal, plant, environmental, fungal, industrial products, and victuals. The types within animals, plants, and fungi followed generally recognized species taxonomy. For example, Diptera is a subclass of Insecta, a subclass of Hexapoda, and so on. The class hierarchy is not

an exhaustive list of every known species but is based on the specific species identified in our isolation data. Due to this feature, the distances along the hierarchy are arbitrary. The high-level classes of the ontology (fungi, plants, and animals) contain a set of sub-classes for parts. For example, pollen is a subclass of flower, which is a subclass of plant parts. The high-level classes defined as environmental, products, and victuals contained relevant subclasses, such as pilsner as a subclass of beer as a subclass of beverage. We exhaustively examined all the isolation environments to build our class hierarchy and relational properties. The lowest level of the hierarchy was the specific isolation environment for each yeast.

The final step, step 5, was to create an instance of each of our yeasts in our hierarchy and assign it to the proper classes and relationships based on the description of its isolation environment. We decided the most specific class would represent the direct environment from which the yeast was isolated. For example, if a yeast was isolated from a beetle on a flower, the beetle was considered the primary or direct class. The association with flowers would be a relational property defined as “is from the animal on the plant.” The ontology contained 1,088 instances (yeasts), 1,569 classes, and 27 object properties. Yeasts with detailed descriptions of their isolation environments were associated with upwards of 20 classes ranging across the hierarchy. Each yeast, however, had only one direct set of classes representing the primary environment (bold red boxes in Figure 1.) Yeasts with sparse descriptions were associated with only a few classes. For example, *Lipomyces tetrasporus* is described only as being isolated from soil. Therefore, its classes are limited to soil-environment, which is a subclass of terrestrial environment, and then environmental classes.

This yeast isolation ontology was developed using Web Protégé (http://protege.stanford.edu), which is a part of the Protégé project (Musen & Protege, 2015). It is presented in the standard Web Ontology Language (OWL) file format for downstream analysis. We have also provided the OWL file for the yeast ontology as a part of the recent publication’s supplement (Opulente et al., 2024).

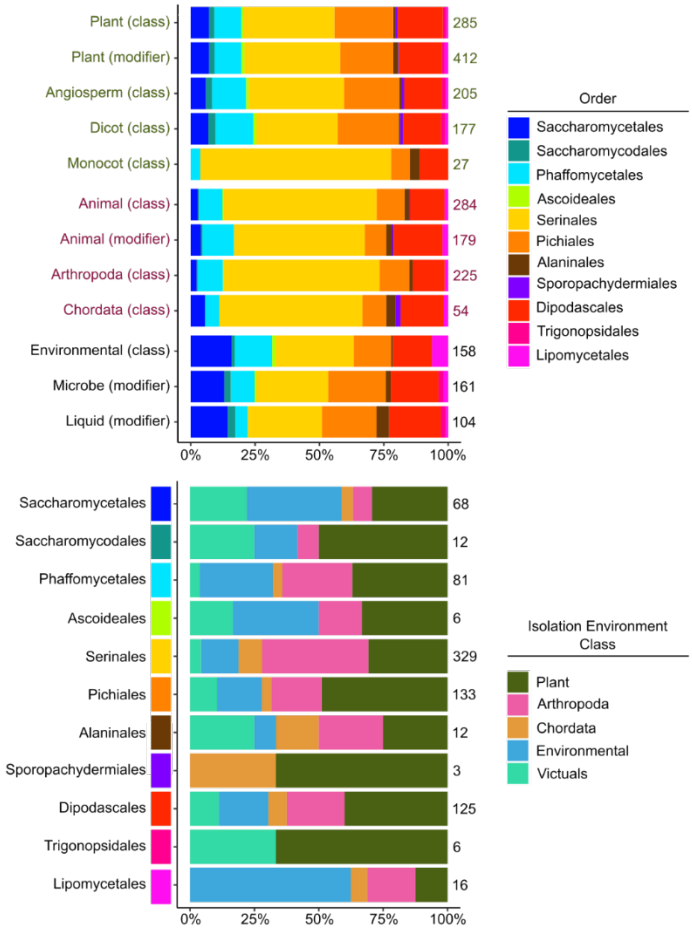


Figure 2: Relative distribution of the isolation environments in the ontology which includes 1,088 yeasts. *A) The categories labeled "class" include yeasts that are an instance of that class or any of its subclasses. The categories labeled "modifier" are those connected to that class by a relationship. For example, any instance that contains the modifier "is from plant on animal" would be included in "Animal (modifier)." These classes are not exclusive—a yeast can be counted in both the "Plant" and "Angiosperm" categories. B) Each order is divided into one of 5 exclusive categories, which are all classes. Therefore, no yeast is counted twice in this section. Not all yeasts, however, are classified into these groups. For example, there are 430 Serinales in this dataset; due to the small overall number of samples, those sampled from other fungi are not shown.*

Interrogation of Yeast Ecology using the Ontology of Yeast Environments (OYE)

The ontology allows us to interrogate where the 1,088 yeasts were isolated from. For example, we saw a higher proportion of Pichiales yeasts in classes associated with the plants class (65/285: 23%) than with the animals class (31/284: 11% Figure 2A). We also interrogated which environments were predominant within each recently established yeast order (Groenewald et al., 2023). The majority of Lipomycetales yeasts were isolated from the environment class (10/16: 63%), and almost half of the Serinales were isolated from the Arthropoda class

(136/329: 41%; Figure 2B). The ontology can also be interrogated at much more refined levels. There were 124 classes that contained between five and ten instances. There were five yeasts that were isolated from mushroom fruiting bodies: *Candida inulinophila*, *Candida morakotiae*, *Candida smagusa*, *Kodamaea fukazawae*, and *Kodamaea fungicola*, which all belong to the order Serinales. There were 6 yeasts isolated from cows: *Nakazawaea peltata* (Alaninales), *Kockiozyma suomiensis* (Lipomycetales), *Wickerhamomyces bovis* (Phaffomycetales), *Magnusiomyces capitatus, Yarrowia hollandica,* and *Zygoascus hellenicus* (Dipodascales).

Classification of yeasts isolated from plants and animals using genomic data

To further demonstrate the utility of the ontology, we conducted a machine learning analysis aimed at identifying genes or pathways associated with specific classes in our ontology. We trained a random forest algorithm using the R programming language to classify yeasts as present or absent in each of the ontology classes (Figure 3A.) The binary data matrix generated from the ontology differentiated between direct sub-classifications and the relational values between (black lines versus colored lines in Figure 1.) The features used to train the model were the presence or absence of genes identified by the Kyoto Encyclopedia of Genes and Genomes (KEGG), which was previously generated across all 1,154 yeasts (Opulente et al., 2024). Briefly, the random forest parameters were tuned to maximize the accuracy and precision of the model. These parameters were then used to train a random forest model using a balanced dataset where 20% of the data was withheld for testing, and 80% was used for training. Models that classified yeasts better than random were then further interrogated by repeating the random forest construction 100 times to examine the impact of the training dataset. The code and complete results can be found in the FigShare repository.

The two most successful models were able to classify yeasts into the class plant (mean AUC of 0.67) or class animal (mean AUC of 0.71). AUC is the area under the receiver operating characteristic curve (ROC), which compares accuracy and precision. Accuracy is a measure of the overall classification success and precision is a measure of per-class success. Therefore, we can classify the isolation environments of yeasts isolated from plants and animals much better than random from gene presence/absence data. The success of these two specific categories is likely related to their large sample sizes (366 for plant and 339 for animal). We then investigated the yeasts that were consistently misclassified (false positives and false negatives) by the algorithm (FigShare Repository.) We noted that a substantial number of yeasts (284 yeasts) were falsely classified as belonging to the plant class if they were isolated from insects associated with plants. Additionally, many yeasts (109) isolated from decaying or dead plants were falsely classified as

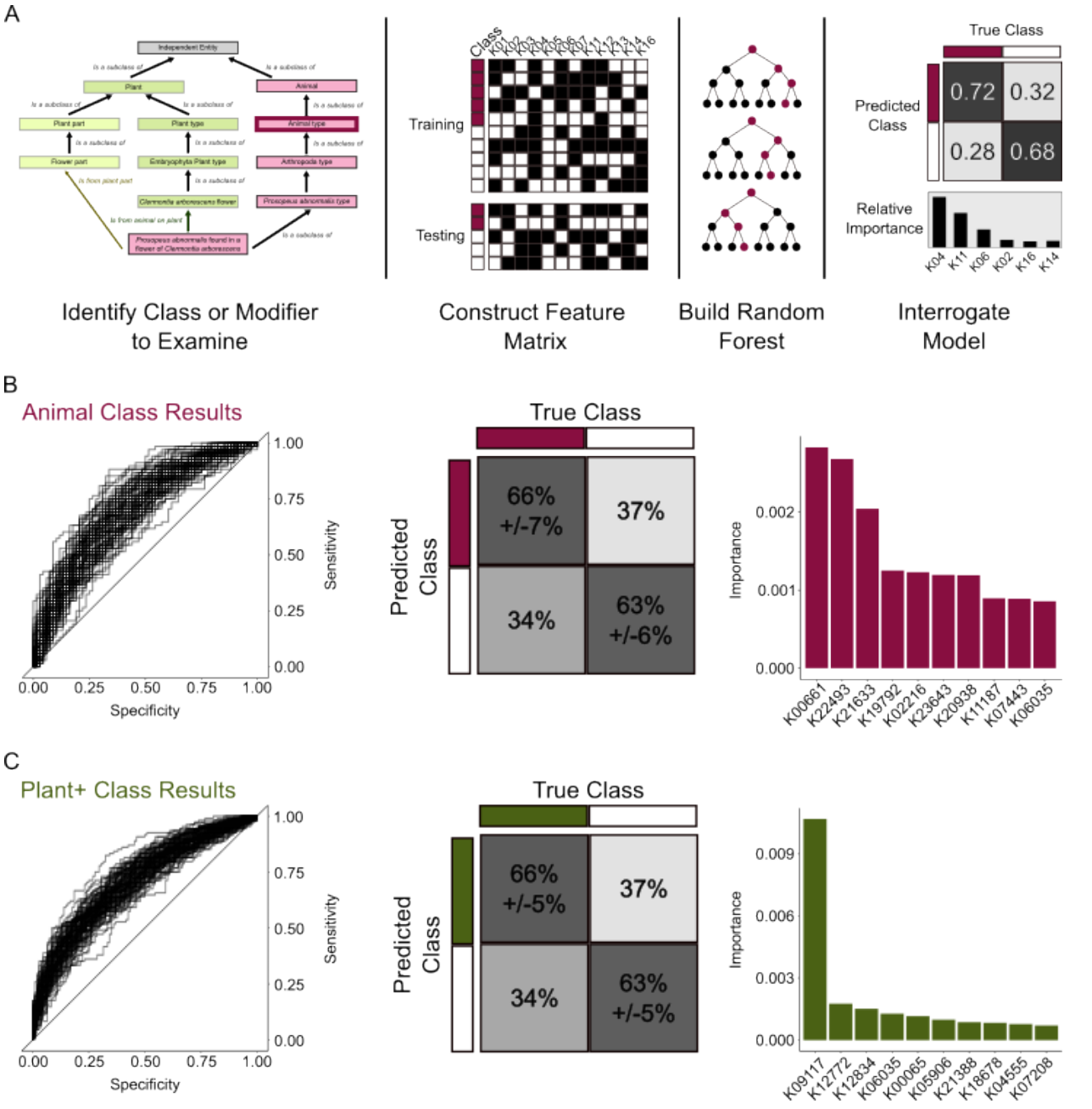


Figure 3 – The Ontology of Yeast Environments enabled machine learning analysis identify genes associated with specific environments. A) The general framework for utilizing the yeast ecological ontology for machine learning. We identified a specific class of interested and obtained all the instances (yeast strains) either directly (black arrows) or relationally (colored arrows) associated with that class. The instances were then divided into training and testing datasets where the presence and absence of KEGG Orthologs (KOs) were used as features. We constructed a random forest and then interrogated the model for accuracy and the important features. B) Classification of yeast in the animal class had an average AUC of 0.71 and an average true-positive rate of 66% across 100 iterations of the model. The KOs with the highest permutation importance are shown in the bar graph. C) Classification of yeast in the plant class (including relational associated but with decayed plants removed) had an average AUC of 0.71 and an average true positive rate of 66% across 100 iterations of the model. There was a single KO (K09117) that had three times higher importance as the next most important KO

not associated with plants. We, therefore, reconstructed the model to classify yeasts as belonging to a plant class or having the relational value “from plant” but not the relational value “decayed microbe association.” For example, in the original model, *Metschnikowia shivogae*, which was isolated from “insects of morning glories,” was not included as an instance of plants. Due to the secondary association, *M. shivogae* was changed to a positive instance in the new model. Conversely, *Sugiyamaella lignohabitans*, which was isolated from “decayed wood” was initially included as a positive instance of plants but was subsequently changed to a negative instance due to its association with decay. When these adjustments were made, the model performance improved from a mean AUC of 0.67 to 0.71. Using the ontology allowed us to easily adjust our data to capture various aspects of the association between yeasts and plants.

Genes and pathways enriched in animal-associated yeasts

We interrogated the KEGG genes that had the highest median permutation importance across the iterations of the models. This analysis allowed us to ask which genes or pathways are important for classifying yeasts as associated with animals. In every model iteration, the KEGG ortholog (KO) K00661 was in the top 1,000 most important features and had the highest median importance (0.0028). This KO encodes a maltose O-acetyltransferase and is annotated in the *S. cerevisiae* genome as an uncharacterized ORF with the systematic name YJL218W. In yeasts isolated from animals, 87% (295/339) have a copy of this gene compared to only 66% (495/747) of non-animal yeasts. Previous work has shown that Oaf1p/Pip2p induces this gene in *S. cerevisiae* in the presence of oleate (Smith et al., 2002). In turn, these regulatory genes (*OAF1/PIP2*) are required for peroxisome proliferation in response to oleate, and their deletion prevents the use of oleate as a singular carbon source (Rottensteiner et al., 2003). Moreover, the YJL218W deletion strain of *S. cerevisiae* had decreased cell membrane integrity and reduced capacity to grow in high salt concentrations (Li et al., 2022). In a general framework, the presence of K00661 may improve yeasts' ability to respond to stressors of the animal environment, especially increased salt concentrations (Manzanares-Estreder et al., 2017). The Na+ salt concentration of insect hemolymph can reach 118 mmol/l (Natochin & Parnova, 1987), while normal human blood sodium levels are ~140 mmol/l (Li et al., 2016). More specifically, the exterior and interior of insects have lipids, including oleic acid, that can both stimulate and prevent fungal growth (Keyhani, 2018). Yeasts associated with insects comprise most of the animal-associated yeasts in our dataset (254/339.) Of the 254 insect-associated yeasts, 227 (89%) have a copy of K00661. In addition to a general role in stress response, genes belonging to the KO K00661 may facilitate growth on and in insects.

We also examined the pathways enriched with genes important for classifying yeasts as animal associated. To identify these pathways, we conducted a KEGG enrichment using the 209 KOs identified in the feature importance analysis of our model. The lysosome pathway (ko04142) was the most highly enriched for KOs identified with our model (seven KOs), although it did not pass statistical significance (adjusted p-value 0.1). This pathway generally corresponds to the function of vacuoles in yeasts as they do not contain lysosomes. The important features of our model had vacuole-associated functions in enzyme transport (K12398, K12397, K12394), acid hydrolases (K12373, K01192, K12350), and membrane proteins (K12386). Two of these KOs

(K12386 and K12394) had a lower abundance in animal-associated yeasts. Animal-associated yeasts are enriched in K12397 and K12398, which are both subunits of the AP-3 complex. K12397 is the β-subunit (Apl6p in *S. cerevisiae*), and K12398 is the μ-subunit (Apm3p in *S. cerevisiae.*) The AP-3 complex is involved in the selective transport of proteins from the Golgi to the vacuole (Cowles et al., 1997). The proteins transported by the AP-3 complex in *S. cerevisiae* are alkaline phosphatases (Cowles et al., 1997), a t-SNARE Vam3p (Cowles et al., 1997), yeast casein kinase 3 (Yck3p) (Sun et al., 2004), and the Niemann-Pick Type C homolog Ncr1p (Berger et al., 2007). Recent work has linked AP-3 with stress-induced vacuole fusion mediated by the protein Yck3p (44, 45) and cell death in both *S. cerevisiae* and the human pathogenic Basidiomycetous yeast *Cryptococcus neoformans* (Stolp et al., 2022). The authors who uncovered the association between AP-3 and cell death suggest that differential regulation of AP-3 may be important for human fungal pathogens (Stolp et al., 2022), which are generally included as animal-associated in our dataset. Interestingly, in our dataset, only K12397 is elevated in human fungal pathogens (10/11) as opposed to their relatives (49/60), according to designations from our previous work (Opulente et al., 2024). Both KOs also have a higher abundance in insect-associated yeasts (50% vs. 40% for K12398 and 82% vs. 77% for K12397).

Three acid hydrolases were also enriched in our models' important features. These are; K12350, a sphingomyelin phosphodiesterase (Ppn1p in *S. cerevisiae*)*;* K12373, a beta-N-hexosaminidase (Hex1p in *C. albicans*)*;* and, K01192, a β-mannosidase (orf19.2838 in *C. albicans*)*.* When transported via vacuoles to the cell exterior, these proteins may break down or modify the environment to allow yeasts to obtain nutrients or combat stressors. Ppn1p cleaves polyphosphates, potentially allowing the use of polyphosphates (45) for protection from oxidative stress (46), formation of canals in the cell wall (47), or as an energy source (Rao et al., 2009). Hex1p is involved in utilizing amino-sugars, such as N-acetyl-D-glucosamine (GlcNAc). In *C. albicans*, this gene is critical for full virulence (Jenkinson & Shepherd, 1987) and plays a role in carbon and nitrogen scavenging during infection of mouse kidneys (Ruhela et al., 2015). Finally, the β-mannosidase has been shown to impact sensitivity to amphotericin B (Xu et al., 2007) and is associated with biofilm production (Bonhomme et al., 2011). We have also found K01192 to be associated with carbon generalism in yeasts (Opulente et al., 2024).

Our analyses illustrate how the availability of a subphylum-wide yeast environment bio-ontology can be employed to identify candidate genes and pathways that may be involved in the adaptation of yeast species to animal environments. Most animal-associated yeasts are associated with arthropods (254/339), while 74 of the remaining yeasts are associated with chordates. Yeasts that are directly associated with animals and arthropod environments experience many of the same stressors, including immune cells, oxidative stress, high salinity, nutrient availability, and even temperature stress as global temperatures rise. The oleate metabolism and vacuole-associated acid hydrolase genes we identified here may be important for the adaptation to these shared stressors.

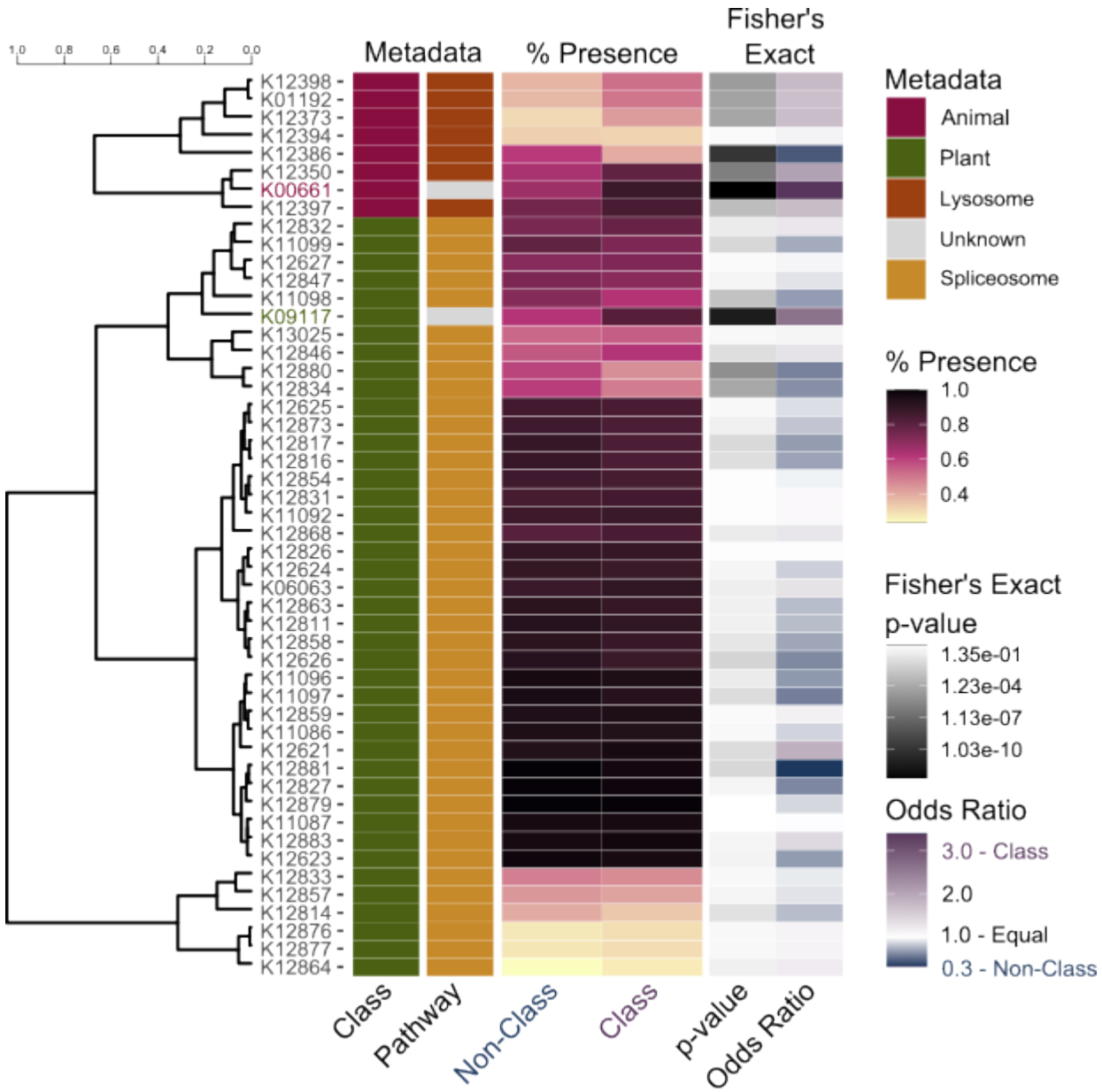


Figure 3: KOs with known and unknown functions were highly informative in the construction of the random forest to classify yeast as isolated from plants or animals. The KOs associated with classification of yeasts in the animal or plant classes (first column) were clustered according to presence in the analyzed class (% Presence columns). The associated pathway for each KO is shown in column 2 with the two most important KOs (colored names) belonging to no known pathway. We also tested for statistical differences in the presence of the KOs in the yeasts belonging to the examined class as compared to those not in that class using a Fisher's exact test. The p-value and odds ratio are reported in the last two columns and the raw data is presented in the FigShare repository.

Genes and pathways enriched in plant-associated yeasts

Finally, we interrogated the model that classified yeasts as plant-associated, including those with secondary modifier associations but no association with decay. The KEGG with the highest

consistent importance in the model was K09117. This is an uncharacterized protein known as Aim41p in *S. cerevisiae*. This gene was found in 83% (364/439) of yeasts associated with plants as compared to 61% (245/404) from non-plant environments. Previous work associated this gene with mitochondrial inheritance (Hess et al., 2009) and upregulation in stress-resistant cells found in the upper level of yeast colonies (Cáp et al., 2012). This gene is over-expressed in *S. cerevisiae* under oxidative stress when exposed to cocoa powder extract (Pelaez-Soto et al., 2020). Other work has shown that the allele-specific expression of *AIM41* is involved in the differential thermal tolerance of *S. cerevisiae* and *S. uvarum* (Li & Fay, 2017). Recently, thermotolerance, but not this specific KEGG, has been implicated in the evolution of yeasts associated with cacti (Goncalves et al., 2023). These results suggest that plant-associated yeasts may be able to better respond to the stressors of the plant environment, such as high temperature due to solar exposure and oxidative stress in plants (Hasanuzzaman & Fujita, 2022).

The spliceosome was the only pathway statistically enriched in the KEGGs important for classifying plant-associated yeasts. 41 KEGGs associated with the spliceosome were also associated with isolation from plants. The KEGG with the highest importance involved in the spliceosome was K12834 (median importance 0.0015), a PHD finger-like domain-containing protein 5A and known as Rds3p in *S. cerevisiae.* This KEGG is absent in 51% (281/550) of the plant-associated yeasts and 39% (211/536) in the non-plant-associated. Despite high conservation in the spliceosome of eukaryotes, previous work in yeasts has shown high variability in the spliceosome, which is likely associated with the loss of introns across the group (Bon et al., 2003). Alterations in the major components of the spliceosome, especially in U4/U5/U6 tri-snRNP, have been shown in yeasts during heat stress response (Bond, 2006; Bracken & Bond, 1999). Two components of the U4/U5/U6 tri-snRNP were important in our model; these were the SM (SNRPB/D2/E/F/G) and LSM (Like Sm; including LSM2/4/5/6/7/8) proteins. We hypothesize, therefore, that the presence and absence of specific spliceosome components may increase or decrease a yeast's ability to respond to specific stressors.

Yeasts associated with the plant or plant-insect environment have a distinct set of important features when compared to animal-associated yeasts. This suggests that the stressors of the plant-insect environment are also distinct. The exact stressors that Aim41p and the spliceosome respond to in the plant environment are not fully elucidated, but both pathways have been associated with heat tolerance.

FUTURE PERSPECTIVES

The ontology of yeast environments (OYE) allowed us to transform individual yeast species descriptions written in natural language into a format interpretable to machine learning algorithms, enabling subphylum-level systematic analyses of yeast isolation environments. By training our machine learning model using gene presence and absence features, we could classify yeasts into those isolated from animals and those isolated from plant or plant-associated environments. Given that yeasts are likely to be found in multiple environments and that adaptation to these environments is likely highly pleiotropic, it is remarkable that our model

reaches an accuracy better than random. In our dataset, we were able to uncover novel associations between genes or pathways and yeasts that were isolated from specific environments.

We anticipate that this ontological framework for isolation environments will be foundational and enable computational complex analysis of wide-ranging yeast ecological data. When DNA is collected from an environment, the metadata often includes natural language descriptors similar to species descriptions. For example, metagenomic samples have recently been collected from soybean rhizosphere (MGNify MGYS00006228) and a whale's blow hole (MGNify MGYS00006536). While natural language interpretation of these environments allows us to know that they are very different, downstream data analysis will require a framework, such as an ontology.

The OYE was created with the explicit purpose of interrogating strain-specific variation in isolation environments associated with the Y1000+ Project genomes (Opulente et al., 2024). To improve the breadth of the ontology, the Y1000+ Project is also adding additional strains for the species sequenced in the. While we believe that this ontology serves as a foundational resource, maintaining and expanding it to capture all of yeast diversity would require a substantial commitment from yeast researchers and culture collections. Therefore, the OYE created here can serve as a model upon which a universal yeast environment ontology could be created. Alternatively, researchers can adapt the OYE to suit their individual needs.

Our ability to connect yeast traits to their environments is only as good as our environmental data. An ontology allows us to capture many aspects of yeast environments in a format that enables the use of powerful machine-learning algorithms. The ontology is also adaptable to historical natural language descriptions and modern metadata collection. Just as phylogenies have enabled investigation of the history of the yeast subphylum, a formalized ontology could transform the way we study the role of environment in yeast function and evolution.

OUTSTANDING QUESTIONS

- The construction of the environmental ontology relies heavily on the natural language descriptions recorded during strain or metagenomic sampling. How can we adapt standards that improve the detail in these descriptions to better capture primary and secondary associations?
- How can we integrate the rapidly growing body of genomic, ecological, and phenotypic data to identify yeast adaptations in response to specific environmental niches?
- Can we integrate the environmental and metagenomic data with our ecological ontology to compare across environments?

ACKNOWLEDGMENTS

We thank Trey K. Sato for feedback. This work was primarily supported by the National Science Foundation (grants DEB-2110403 to C.T.H. and DEB-2110404 to A.R.). Computational analyses were run in the UNC Charlotte high performance computing cluster in Charlotte North

Carolina. X.-X.S. was supported by the NSF 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). Research in the Hittinger Lab is also supported by the USDA National Institute of Food and Agriculture (Hatch Project 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 NIH/National Institute of Allergy and Infectious Diseases (R01 AI153356), and the Burroughs Wellcome Fund.

CONFLICT OF INTEREST STATEMENT

AR is a scientific consultant for LifeMine Therapeutics, Inc. The other authors declare no other competing interests.

DATA SHARING AND DATA AVAILABILITY

The Y1000+ data can be obtained from the project website (http://y1000plus.org). The Figshare repository https://figshare.com/projects/Exploring_Saccharomycotina_Yeast_Ecology_Through_an_Ecological_Ontology_Framework/208648 the raw random forest model data and a copy of the ontology.

REFERENCES

Abzhanov, A., Protas, M., Grant, B. R., Grant, P. R., & Tabin, C. J. (2004). Bmp4 and morphological variation of beaks in Darwin's finches. *Science*, *305*(5689), 1462-1465. https://doi.org/10.1126/science.1098095

Alsammar, H. F., Naseeb, S., Brancia, L. B., Gilman, R. T., Wang, P., & Delneri, D. (2019). Targeted metagenomics approach to capture the biodiversity of Saccharomyces genus in wild environments. *Environ Microbiol Rep*, *11*(2), 206-214. https://doi.org/10.1111/1758-2229.12724

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, *25*(1), 25-29. https://doi.org/10.1038/75556

Berger, A. C., Salazar, G., Styers, M. L., Newell-Litwa, K. A., Werner, E., Maue, R. A., Corbett, A. H., & Faundez, V. (2007). The subcellular localization of the Niemann-Pick Type C proteins depends on the adaptor complex AP-3. *J Cell Sci*, *120*(Pt 20), 3640-3652. https://doi.org/10.1242/jcs.03487

Bidaud, A. L., Chowdhary, A., & Dannaoui, E. (2018). Candida auris: An emerging drug resistant yeast - A mini-review. *J Mycol Med*, *28*(3), 568-573. https://doi.org/10.1016/j.mycmed.2018.06.007

Blackwell, M. (2017). Made for Each Other: Ascomycete Yeasts and Insects. *Microbiol Spectr*, *5*(3). https://doi.org/10.1128/microbiolspec.FUNK-0081-2016

Bon, E., Casaregola, S., Blandin, G., Llorente, B., Neuveglise, C., Munsterkotter, M., Guldener, U., Mewes, H. W., Van Helden, J., Dujon, B., & Gaillardin, C. (2003). Molecular evolution of eukaryotic genomes: hemiascomycetous yeast spliceosomal introns. *Nucleic Acids Res*, *31*(4), 1121-1135. https://doi.org/10.1093/nar/gkg213

589 Bond, U. (2006). Stressed out! Effects of environmental stress on mRNA metabolism. *FEMS Yeast Res*,
590 *6*(2), 160-170. https://doi.org/10.1111/j.1567-1364.2006.00032.x
591 Bonhomme, J., Chauvel, M., Goyard, S., Roux, P., Rossignol, T., & d'Enfert, C. (2011). Contribution of the
592 glycolytic flux and hypoxia adaptation to efficient biofilm formation by Candida albicans. *Mol*
593 *Microbiol*, *80*(4), 995-1013. https://doi.org/10.1111/j.1365-2958.2011.07626.x
594 Botha, A. (2011). The importance and ecology of yeasts in soil. *Soil Biology and Biochemistry*, *43*(1), 1-8.
595 Bowles, J. M., & Lachance, M. A. (1983). Patterns of Variation in the Yeast Florae of Exudates in an Oak
596 Community. *Canadian Journal of Botany-Revue Canadienne De Botanique*, *61*(12), 2984-2995.
597 https://doi.org/DOI 10.1139/b83-335
598 Bracken, A. P., & Bond, U. (1999). Reassembly and protection of small nuclear ribonucleoprotein
599 particles by heat shock proteins in yeast cells. *Rna*, *5*(12), 1586-1596.
600 https://doi.org/10.1017/s1355838299991203
601 Brejova, B., Lichancova, H., Brazdovic, F., Hegedusova, E., Forgacova Jakubkova, M., Hodorova, V.,
602 Dzugasova, V., Balaz, A., Zeiselova, L., Cillingova, A., Nebohacova, M., Raclavsky, V., Tomaska, L.,
603 Lang, B. F., Vinar, T., & Nosek, J. (2019). Genome sequence of the opportunistic human
604 pathogen Magnusiomyces capitatus. *Curr Genet*, *65*(2), 539-560.
605 https://doi.org/10.1007/s00294-018-0904-y
606 Brettner, L., Ho, W. C., Schmidlin, K., Apodaca, S., Eder, R., & Geiler-Samerotte, K. (2022). Challenges and
607 potential solutions for studying the genetic and phenotypic architecture of adaptation in
608 microbes. *Curr Opin Genet Dev*, *75*, 101951. https://doi.org/10.1016/j.gde.2022.101951
609 Butinar, L., Strmole, T., & Gunde-Cimerman, N. (2011). Relative incidence of ascomycetous yeasts in
610 arctic coastal environments. *Microb Ecol*, *61*(4), 832-843. https://doi.org/10.1007/s00248-010-
611 9794-3
612 Buttigieg, P. L., Morrison, N., Smith, B., Mungall, C. J., Lewis, S. E., & Consortium, E. (2013). The
613 environment ontology: contextualising biological and biomedical entities. *J Biomed Semantics*,
614 *4*(1), 43. https://doi.org/10.1186/2041-1480-4-43
615 Cadete, R. M., Lopes, M. R., & Rosa, C. A. (2017). Yeasts associated with decomposing plant material and
616 rotting wood. *Yeasts in natural ecosystems: diversity*, 265-292.
617 Cáp, M., Stepánek, L., Harant, K., Váchová, L., & Palková, Z. (2012). Cell Differentiation within a Yeast
618 Colony: Metabolic and Regulatory Parallels with a Tumor-Affected Organism. *Molecular Cell*,
619 *46*(4), 436-448. https://doi.org/10.1016/j.molcel.2012.04.001
620 Cavalieri, D., Valentini, B., & Stefanini, I. (2022). Going wild: ecology and genomics are crucial to
621 understand yeast evolution. *Curr Opin Genet Dev*, *75*, 101922.
622 https://doi.org/10.1016/j.gde.2022.101922
623 Cowles, C. R., Odorizzi, G., Payne, G. S., & Emr, S. D. (1997). The AP-3 adaptor complex is essential for
624 cargo-selective transport to the yeast vacuole. *Cell*, *91*(1), 109-118. https://doi.org/Doi
625 10.1016/S0092-8674(01)80013-1
626 Cunha, A. O. B., Bezerra, J. D. P., Oliveira, T. G. L., Barbier, E., Bernard, E., Machado, A. R., & Souza-
627 Motta, C. M. (2020). Living in the dark: Bat caves as hotspots of fungal diversity. *PLoS One*,
628 *15*(12), e0243494. https://doi.org/10.1371/journal.pone.0243494
629 Dahdul, W., Balhoff, J., Lapp, H., Uyeda, J., & Vision, T. (2017). *Enabling machine-actionable semantics*
630 *for comparative analyses of trait evolution* [Grant].
631 David, K. T., Harrison, M. C., Opulente, D. A., LaBella, A. L., Wolters, J. F., Zhou, X., Shen, X. X.,
632 Groenewald, M., Pennell, M., Hittinger, C. T., & Rokas, A. (2024). Saccharomycotina yeasts defy
633 long-standing macroecological patterns. *Proc Natl Acad Sci U S A*, *121*(10), e2316031121.
634 https://doi.org/10.1073/pnas.2316031121
635 Dooley, D. M., Griffiths, E. J., Gosal, G. S., Buttigieg, P. L., Hoehndorf, R., Lange, M. C., Schriml, L. M.,
636 Brinkman, F. S. L., & Hsiao, W. W. L. (2018). FoodOn: a harmonized food ontology to increase

global food traceability, quality control and data integration. *Npj Science of Food*, *2*(1). https://doi.org/ARTN 23

10.1038/s41538-018-0032-6

Edmunds, R. C., Su, B., Balhoff, J. P., Eames, B. F., Dahdul, W. M., Lapp, H., Lundberg, J. G., Vision, T. J., Dunham, R. A., Mabee, P. M., & Westerfield, M. (2016). Phenoscape: Identifying Candidate Genes for Evolutionary Phenotypes. *Mol Biol Evol*, *33*(1), 13-24. https://doi.org/10.1093/molbev/msv223

GBIF: The Global Biodiversity Information Facility (2024). https://www.gbif.org

Goncalves, C., Harrison, M. C., Steenwyk, J. L., Opulente, D. A., LaBella, A. L., Wolters, J. F., Zhou, X., Shen, X. X., Groenewald, M., Hittinger, C. T., & Rokas, A. (2023). Diverse signatures of convergent evolution in cacti-associated yeasts. *bioRxiv*. https://doi.org/10.1101/2023.09.14.557833

Groenewald, M., Hittinger, C., Bensch, K., Opulente, D., Shen, X.-X., Li, Y., Liu, C., LaBella, A., Zhou, X., & Limtong, S. (2023). A genome-informed higher rank classification of the biotechnologically important fungal subphylum Saccharomycotina. *Studies in Mycology*.

Haendel, M. A., Neuhaus, F., Osumi-Sutherland, D., Mabee, P. M., Mejino Jr, J. L., Mungall, C. J., & Smith, B. (2008). CARO–the common anatomy reference ontology. In *Anatomy ontologies for bioinformatics: principles and practice* (pp. 327-349). Springer.

Hagman, A., & Piskur, J. (2015). A study on the fundamental mechanism and the evolutionary driving forces behind aerobic fermentation in yeast. *PLoS One*, *10*(1), e0116942. https://doi.org/10.1371/journal.pone.0116942

Harrison, M. C., Ubbelohde, E. J., LaBella, A. L., Opulente, D. A., Wolters, J. F., Zhou, X., Shen, X. X., Groenewald, M., Hittinger, C. T., & Rokas, A. (2024). Machine learning enables identification of an alternative yeast galactose utilization pathway. *Proc Natl Acad Sci U S A*, *121*(18), e2315314121. https://doi.org/10.1073/pnas.2315314121

Hasanuzzaman, M., & Fujita, M. (2022). Plant Oxidative Stress: Biology, Physiology and Mitigation. *Plants (Basel)*, *11*(9). https://doi.org/10.3390/plants11091185

Hastings, J. (2017). Primer on Ontologies. *Methods Mol Biol*, *1446*, 3-13. https://doi.org/10.1007/978-1-4939-3743-1_1

Hess, D. C., Myers, C. L., Huttenhower, C., Hibbs, M. A., Hayes, A. P., Paw, J., Clore, J. J., Mendoza, R. M., Luis, B. S., Nislow, C., Giaever, G., Costanzo, M., Troyanskaya, O. G., & Caudy, A. A. (2009). Computationally driven, quantitative experiments discover genes required for mitochondrial biogenesis. *PLoS Genet*, *5*(3), e1000407. https://doi.org/10.1371/journal.pgen.1000407

Hittinger, C. T., Goncalves, P., Sampaio, J. P., Dover, J., Johnston, M., & Rokas, A. (2010). Remarkably ancient balanced polymorphisms in a multi-locus gene network. *Nature*, *464*(7285), 54-58. https://doi.org/10.1038/nature08791

Hittinger, C. T., Steele, J. L., & Ryder, D. S. (2018). Diverse yeasts for diverse fermented beverages and foods. *Curr Opin Biotechnol*, *49*, 199-206. https://doi.org/10.1016/j.copbio.2017.10.004

Jaiswal, P., Avraham, S., Ilic, K., Kellogg, E. A., McCouch, S., Pujar, A., Reiser, L., Rhee, S. Y., Sachs, M. M., Schaeffer, M., Stein, L., Stevens, P., Vincent, L., Ware, D., & Zapata, F. (2005). Plant Ontology (PO): a Controlled Vocabulary of Plant Structures and Growth Stages. *Comp Funct Genomics*, *6*(7-8), 388-397. https://doi.org/10.1002/cfg.496

Jenkinson, H. F., & Shepherd, M. G. (1987). A mutant of Candida albicans deficient in beta-N-acetylglucosaminidase (chitobiase). *J Gen Microbiol*, *133*(8), 2097-2106. https://doi.org/10.1099/00221287-133-8-2097

Keeler, E., Burgaud, G., Teske, A., Beaudoin, D., Mehiri, M., Dayras, M., Cassand, J., & Edgcomb, V. (2021). Deep-sea hydrothermal vent sediments reveal diverse fungi with antibacterial activities. *FEMS Microbiol Ecol*, *97*(8). https://doi.org/10.1093/femsec/fiab103

Keyhani, N. O. (2018). Lipid biology in fungal stress and virulence: Entomopathogenic fungi. *Fungal Biol*, *122*(6), 420-429. https://doi.org/10.1016/j.funbio.2017.07.003

Kuhn, M., & Vaughan, D. (2024). *parsnip: A Common API to Modeling and Analysis Functions*. In (Version R package version 1.2.1) https://parsnip.tidymodels.org/

Kurtzman, C., Fell, J. W., & Boekhout, T. (2011). *The yeasts: a taxonomic study*. Elsevier.

LaBella, A. L., Opulente, D. A., Steenwyk, J. L., Hittinger, C. T., & Rokas, A. (2021). Signatures of optimal codon usage in metabolic genes inform budding yeast ecology. *PLoS Biol*, *19*(4), e3001185. https://doi.org/10.1371/journal.pbio.3001185

Lachance, M. A. (2020). Guidelines for the publication of novel yeast species descriptions in Yeast. *Yeast*, *37*(3), 251-252. https://doi.org/10.1002/yea.3465

Lee, K. B., Wang, J., Palme, J., Escalante-Chong, R., Hua, B., & Springer, M. (2017). Polymorphisms in the yeast galactose sensor underlie a natural continuum of nutrient-decision phenotypes. *PLoS Genet*, *13*(5), e1006766. https://doi.org/10.1371/journal.pgen.1006766

Levy, R., & Borenstein, E. (2012). Reverse Ecology: from systems to environments and back. *Adv Exp Med Biol*, *751*, 329-345. https://doi.org/10.1007/978-1-4614-3567-9_15

Li, H., Sun, S. R., Yap, J. Q., Chen, J. H., & Qian, Q. (2016). 0.9% saline is neither normal nor physiological. *J Zhejiang Univ Sci B*, *17*(3), 181-187. https://doi.org/10.1631/jzus.B1500201

Li, M., Zhang, Y., Deng, J., Wang, H., Ma, J., Wang, W., & Lyu, L. (2022). Deletion of YJL218W reduces salt tolerance of Saccharomyces cerevisiae. *J Basic Microbiol*, *62*(8), 930-936. https://doi.org/10.1002/jobm.202200029

Li, X. Y. C., & Fay, J. C. (2017). Regulatory Divergence in Gene Expression between Two Thermally Divergent Yeast Species. *Genome Biology and Evolution*, *9*(5), 1120-1129. https://doi.org/10.1093/gbe/evx072

Loureiro, V., & Querol, A. (1999). The prevalence and control of spoilage yeasts in foods and beverages. *Trends in Food Science & Technology*, *10*(11), 356-365. https://doi.org/Doi 10.1016/S0924-2244(00)00021-2

Manzanares-Estreder, S., Espí-Bardisa, J., Alarcón, B., Pascual-Ahuir, A., & Proft, M. (2017). Multilayered control of peroxisomal activity upon salt stress in. *Molecular Microbiology*, *104*(5), 851-868. https://doi.org/10.1111/mmi.13669

Morais, C. G., Cadete, R. M., Uetanabaro, A. P., Rosa, L. H., Lachance, M. A., & Rosa, C. A. (2013). D-xylose-fermenting and xylanase-producing yeast species from rotting wood of two Atlantic Rainforest habitats in Brazil. *Fungal Genetics and Biology*, *60*, 19-28. https://doi.org/10.1016/j.fgb.2013.07.003

Musen, M. A., & Protege, T. (2015). The Protege Project: A Look Back and a Look Forward. *AI Matters*, *1*(4), 4-12. https://doi.org/10.1145/2757001.2757003

Nagahama, T. (2006). Yeast biodiversity in freshwater, marine and deep-sea environments. In *Biodiversity and ecophysiology of yeasts* (pp. 241-262). Springer.

Nagano, Y., Miura, T., Tsubouchi, T., Lima, A. O., Kawato, M., Fujiwara, Y., & Fujikura, K. (2020). Cryptic fungal diversity revealed in deep-sea sediments associated with whale-fall chemosynthetic ecosystems. *Mycology*, *11*(3), 263-278. https://doi.org/10.1080/21501203.2020.1799879

Nagano, Y., Nagahama, T., & Abe, F. (2014). Cold-adapted yeasts in deep-sea environments. *Cold-adapted Yeasts: Biodiversity, Adaptation Strategies and Biotechnological Significance*, 149-171.

Nalabothu, R. L., Fisher, K. J., LaBella, A. L., Meyer, T. A., Opulente, D. A., Wolters, J. F., Rokas, A., & Hittinger, C. T. (2023). Codon optimization improves the prediction of xylose metabolism from gene content in budding yeasts. *Mol Biol Evol*. https://doi.org/10.1093/molbev/msad111

Narunsky-Haziza, L., Sepich-Poore, G. D., Livyatan, I., Asraf, O., Martino, C., Nejman, D., Gavert, N., Stajich, J. E., Amit, G., Gonzalez, A., Wandro, S., Perry, G., Ariel, R., Meltser, A., Shaffer, J. P., Zhu, Q., Balint-Lahat, N., Barshack, I., Dadiani, M., . . . Straussman, R. (2022). Pan-cancer analyses reveal cancer-type-specific fungal ecologies and bacteriome interactions. *Cell*, *185*(20), 3789-3806 e3717. https://doi.org/10.1016/j.cell.2022.09.005

Natochin, Y. V., & Parnova, R. G. (1987). Osmolality and Electrolyte Concentration of Hemolymph and the Problem of Ion and Volume Regulation of Cells in Higher Insects. *Comparative Biochemistry and Physiology a-Physiology*, *88*(3), 563-570. https://doi.org/Doi 10.1016/0300-9629(87)90082-X

Noy, N. F., & McGuinness, D. L. (2001). Ontology development 101: A guide to creating your first ontology. In: Stanford knowledge systems laboratory technical report KSL-01-05 and ….

Opulente, D. A., LaBella, A. L., Harrison, M. C., Wolters, J. F., Liu, C., Li, Y., Kominek, J., Steenwyk, J. L., Stoneman, H. R., VanDenAvond, J., Miller, C. R., Langdon, Q. K., Silva, M., Goncalves, C., Ubbelohde, E. J., Li, Y., Buh, K. V., Jarzyna, M., Haase, M. A. B., . . . Hittinger, C. T. (2024). Genomic factors shape carbon and nitrogen metabolic niche breadth across Saccharomycotina yeasts. *Science*, *384*(6694), eadj4503. https://doi.org/10.1126/science.adj4503

Opulente, D. A., Rollinson, E. J., Bernick-Roehr, C., Hulfachor, A. B., Rokas, A., Kurtzman, C. P., & Hittinger, C. T. (2018). Factors driving metabolic diversity in the budding yeast subphylum. *BMC Biol*, *16*(1), 26. https://doi.org/10.1186/s12915-018-0498-3

Pelaez-Soto, A., Roig, P., Martinez-Culebras, P. V., Fernandez-Espinar, M. T., & Gil, J. V. (2020). Proteomic Analysis of Saccharomyces cerevisiae Response to Oxidative Stress Mediated by Cocoa Polyphenols Extract. *Molecules*, *25*(3). https://doi.org/10.3390/molecules25030452

Perez, J. C. (2021). The interplay between gut bacteria and the yeast Candida albicans. *Gut Microbes*, *13*(1), 1979877. https://doi.org/10.1080/19490976.2021.1979877

Peter, J., De Chiara, M., Friedrich, A., Yue, J. X., Pflieger, D., Bergstrom, A., Sigwalt, A., Barre, B., Freel, K., Llored, A., Cruaud, C., Labadie, K., Aury, J. M., Istace, B., Lebrigand, K., Barbry, P., Engelen, S., Lemainque, A., Wincker, P., . . . Schacherer, J. (2018). Genome evolution across 1,011 Saccharomyces cerevisiae isolates. *Nature*, *556*(7701), 339-344. https://doi.org/10.1038/s41586-018-0030-5

Pontes, A., Paraiso, F., Liu, Y. C., Limtong, S., Jindamorakot, S., Jespersen, L., Goncalves, C., Rosa, C. A., Tsai, I. J., Rokas, A., Hittinger, C. T., Goncalves, P., & Sampaio, J. P. (2024). Tracking alternative versions of the galactose gene network in the genus Saccharomyces and their expansion after domestication. *iScience*, *27*(2), 108987. https://doi.org/10.1016/j.isci.2024.108987

Postma, E., Verduyn, C., Scheffers, W. A., & Van Dijken, J. P. (1989). Enzymic analysis of the crabtree effect in glucose-limited chemostat cultures of Saccharomyces cerevisiae. *Appl Environ Microbiol*, *55*(2), 468-477. https://doi.org/10.1128/aem.55.2.468-477.1989

Rao, N. N., Gomez-Garcia, M. R., & Kornberg, A. (2009). Inorganic polyphosphate: essential for growth and survival. *Annu Rev Biochem*, *78*, 605-647. https://doi.org/10.1146/annurev.biochem.77.083007.093039

Riley, R., Haridas, S., Wolfe, K. H., Lopes, M. R., Hittinger, C. T., Goker, 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., . . . Jeffries, T. W. (2016). Comparative genomics of biotechnologically important yeasts. *Proc Natl Acad Sci U S A*, *113*(35), 9882-9887. https://doi.org/10.1073/pnas.1603941113

Rosa, C. A., Morais, P. B., Lachance, M.-A., Pimenta, R. S., Santos, R. O., Trindade, R. C., & Figueroa, D. L. (2006). Candida azymoides sp. n., a yeast species from tropical fruits and larva (Ascomycota) of Anastrepha mucronota (Diptera: Tephritidae). *Lundiana: International Journal of Biodiversity*, *7*(2), 83-86.

Rosenbach, A., Dignard, D., Pierce, J. V., Whiteway, M., & Kumamoto, C. A. (2010). Adaptations of Candida albicans for growth in the mammalian intestinal tract. *Eukaryot Cell*, *9*(7), 1075-1086. https://doi.org/10.1128/EC.00034-10

Rottensteiner, H., Wabnegger, L., Erdmann, R., Hamilton, B., Ruis, H., Hartig, A., & Gurvitz, A. (2003). *Saccharomyces cerevisiae* PIP2 mediating oleic acid induction and peroxisome proliferation is regulated by Adr1p and Pip2p-Oaf1p. *Journal of Biological Chemistry*, *278*(30), 27605-27611. https://doi.org/10.1074/jbc.M304097200

Ruhela, D., Kamthan, M., Saha, P., Majumdar, S. S., Datta, K., Abdin, M. Z., & Datta, A. (2015). In vivo role of Candida albicans beta-hexosaminidase (HEX1) in carbon scavenging. *Microbiologyopen*, *4*(5), 730-742. https://doi.org/10.1002/mbo3.274

Sarabia, M., Cornejo, P., Azcón, R., Carreón-Abud, Y., & Larsen, J. (2017). Mineral phosphorus fertilization modulates interactions between maize, rhizosphere yeasts and arbuscular mycorrhizal fungi. *Rhizosphere*, *4*, 89-93. https://doi.org/10.1016/j.rhisph.2017.09.001

Selbmann, L., Zucconi, L., Onofri, S., Cecchini, C., Isola, D., Turchetti, B., & Buzzini, P. (2014). Taxonomic and phenotypic characterization of yeasts isolated from worldwide cold rock-associated habitats. *Fungal Biol*, *118*(1), 61-71. https://doi.org/10.1016/j.funbio.2013.11.002

Shen, X. X., Opulente, D. A., Kominek, J., Zhou, X., Steenwyk, J. L., Buh, K. V., Haase, M. A. B., Wisecaver, J. H., Wang, M., Doering, D. T., Boudouris, J. T., Schneider, R. M., Langdon, Q. K., Ohkuma, M., Endoh, R., Takashima, M., Manabe, R., Cadez, N., Libkind, D., . . . Rokas, A. (2018). Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum. *Cell*, *175*(6), 1533-+. https://doi.org/10.1016/j.cell.2018.10.023

Shen, X. X., Steenwyk, J. L., LaBella, A. L., Opulente, D. A., Zhou, X., Kominek, J., Li, Y., Groenewald, M., Hittinger, C. T., & Rokas, A. (2020). Genome-scale phylogeny and contrasting modes of genome evolution in the fungal phylum Ascomycota. *Sci Adv*, *6*(45). https://doi.org/10.1126/sciadv.abd0079

Slavikova, E., Vadkertiova, R., & Vranova, D. (2007). Yeasts colonizing the leaf surfaces. *J Basic Microbiol*, *47*(4), 344-350. https://doi.org/10.1002/jobm.200710310

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., Consortium, O. B. I., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S. A., Scheuermann, R. H., Shah, N., Whetzel, P. L., & Lewis, S. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol*, *25*(11), 1251-1255. https://doi.org/10.1038/nbt1346

Smith, J. J., Marelli, M., Christmas, R. H., Vizeacoumar, F. J., Dilworth, D. J., Ideker, T., Galitski, T., Dimitrov, K., Rachubinski, R. A., & Aitchison, J. D. (2002). Transcriptome profiling to identify genes involved in peroxisome assembly and function. *J Cell Biol*, *158*(2), 259-271. https://doi.org/10.1083/jcb.200204059

Starmer, W. T., & Fogleman, J. C. (1986). Coadaptation ofDrosophila and yeasts in their natural habitat. *J Chem Ecol*, *12*(5), 1037-1055. https://doi.org/10.1007/BF01638995

Starmer, W. T., & Lachance, M.-A. (2011). Yeast Ecology. In C. P. F. Kurtzman, Jack W.; Boekhout, Teun (Ed.), *The Yeasts* (Vol. Volume 1, pp. 65-86). Elsevier.

Stefanini, I. (2018). Yeast-insect associations: It takes guts. *Yeast*, *35*(4), 315-330. https://doi.org/10.1002/yea.3309

Stolp, Z. D., Kulkarni, M., Liu, Y., Zhu, C., Jalisi, A., Lin, S., Casadevall, A., Cunningham, K. W., Pineda, F. J., Teng, X., & Hardwick, J. M. (2022). Yeast cell death pathway requiring AP-3 vesicle trafficking leads to vacuole/lysosome membrane permeabilization. *Cell Rep*, *39*(2), 110647. https://doi.org/10.1016/j.celrep.2022.110647

Suhr, M. J., & Hallen-Adams, H. E. (2015). The human gut mycobiome: pitfalls and potentials-a mycologist's perspective. *Mycologia*, *107*(6), 1057-1073. https://doi.org/10.3852/15-147

Sun, B. M., Chen, L. Y., Cao, W., Roth, A. F., & Davis, N. G. (2004). The yeast casein kinase Yck3p is palmitoylated, then sorted to the vacuolar membrane with AP-3-dependent recognition of a YXXΦ adaptin sorting signal. *Molecular Biology of the Cell*, *15*(3), 1397-1406. https://doi.org/10.1091/mbc.E03-09-0682

Vaïtilingom, M., Attard, E., Gaiani, N., Sancelme, M., Deguillaume, L., Flossmann, A. I., Amato, P., & Delort, A.-M. (2012). Long-term features of cloud microbiology at the puy de Dôme (France). *Atmospheric environment*, *56*, 88-100.

Van Slyke, C. E., Bradford, Y. M., Westerfield, M., & Haendel, M. A. (2014). The zebrafish anatomy and stage ontologies: representing the anatomy and development of Danio rerio. *J Biomed Semantics*, *5*(1), 12. https://doi.org/10.1186/2041-1480-5-12

Vetrovsky, T., Morais, D., Kohout, P., Lepinay, C., Algora, C., Awokunle Holla, S., Bahnmann, B. D., Bilohneda, K., Brabcova, V., D'Alo, F., Human, Z. R., Jomura, M., Kolarik, M., Kvasnickova, J., Llado, S., Lopez-Mondejar, R., Martinovic, T., Masinova, T., Meszarosova, L., . . . Baldrian, P. (2020). GlobalFungi, a global database of fungal occurrences from high-throughput-sequencing metabarcoding studies. *Sci Data*, *7*(1), 228. https://doi.org/10.1038/s41597-020-0567-7

Wright, M. N., & Ziegler, A. (2015). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *arXiv preprint arXiv:1508.04409*.

Xu, D., Jiang, B., Ketela, T., Lemieux, S., Veillette, K., Martel, N., Davison, J., Sillaots, S., Trosok, S., Bachewich, C., Bussey, H., Youngman, P., & Roemer, T. (2007). Genome-wide fitness test and mechanism-of-action studies of inhibitory compounds in Candida albicans. *Plos Pathogens*, *3*(6), e92. https://doi.org/10.1371/journal.ppat.0030092

Yu, G., Wang, L.-G., Han, Y., & He, Q.-Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology*, *16*(5), 284-287.

SUPPLEMENTAL METHODS

Random Forest Construction

Random forest construction was conducted in R v4.2.2-mpi. The features used on model construction were the presence and absence (encoded as 0 and 1) of KOs annotated by KEGG obtained from previous work (Opulente et al., 2024). KEGGs with a presence below 20% across all species were removed. An initial random forest was tuned twice using the ranger package v0.16.0 (Wright & Ziegler, 2015) and parsnip v1.2.1 (Kuhn & Vaughan, 2024), withholding 20% of the data for validation. The first tuning was a grid search based on an initial tuning of the model. The mtry (number of variables to split at each node) and min_n (minimum number of data points for node splitting) values obtained from this tuning were then used in another grid search using 0.75 and 1.25 times the values of the first search. The final random forest model parameters were selected based on the model's maximum area under the curve (AUC). We then constructed 100 random forest models using a different training and testing data set for each iteration. For each of the 100 random forest models, we withheld 20% of the data for model construction. The model parameters, classifications, and important features (measured by permutation in the ranger package) were stored for each iteration.

KEGG analysis

It is important to note that we filtered out results from the KEGG pathways labeled " – yeast." Our previous analysis (Opulente et al., 2024) showed that the KEGG database narrowly defines these as pathways in the Saccharomycetales and are under-annotated across species, especially in yeasts from

other orders. We also manually re-checked KEGG presence and absence to verify the results of the automatic KEGG analysis previously conducted and removed KEGGs with significant differences in the re-annotation.

We analyzed KEGG annotations that were identified in the top 1000 most important KEGGs in 80% of the 100 random forest models. These KEGGs were then run through an enrichment analysis to identify enriched pathways. This analysis was conducted in clusterProfiler v 4.10.1 (Yu et al., 2012) using the Benjamini-Hochberg multiple-testing correction. The possible KEGG universe was defined as all the KEGGs annotated in the input yeast genomes. Using a Fisher's exact test, we re-analyzed each KEGG's presence and absence counts across the classifications. We report the raw uncorrected p-value and odds ratio for each KEGG.