Use of Machine Learning to Predict the Desiccation Tolerance of Bacteria

A THESIS
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Maia L C Clipsham

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

Advisors: Dr. Lawrence Wackett

& Dr. Alptekin Aksan

August 2021

Acknowledgements

I would like to thank my advisors, Dr. Lawrence Wackett and Dr. Alptekin Aksan, for all the support and guidance they have given me throughout my graduate studies. I have learned about many things, some of them related to science.

I am grateful for my master's committee, Dr. Michael Smanski and Dr. Kyle Costa for their suggestions and for taking time to discuss science.

I would like to thank Dr. Serina Robinson, for her invaluable knowledge and guidance, without her this project would not have happened.

I would also like to thank my lab members for their valuable discussions and feedback: Beatriz DeSouza, Advitiya Mahajan, Lambros Tassoulas, Megan Smith, Mian Wang, Jack Richardson, Tony Dodge, Maddie Bygd, and Ash Robinson. I would like to particularly thank Kelly Aukema for being an invaluable resource to me with everything in the lab and in Minnesota.

Dedication

To my family and friends for all their support and encouragement throughout my graduate education.

Abstract

For efficient long-term storage and use of bacteria for environmental applications, understanding and identifying desiccation resistance in bacteria is key. In the past, desiccation tolerance was a common way of characterizing bacteria, so there is much data on the desiccation tolerance of a wide range of bacterial species. Since the advent of transcriptomics, multiple papers have been published on the expression level of genes during desiccation stress. Additionally, many reviews have described mechanisms and genes relevant to desiccation tolerance in bacteria, but an overarching framework for the prediction of desiccation survival in bacteria is lacking. Model building based on data collected from the literature has been used to successfully predict aerobic vs anaerobic phenotype, enzyme function and substrate specificity (Robinson et al., 2020; Jabłońska et al, 2019) Building on this wealth of previous research, machine learning was used to create a robust model that predicts desiccation tolerance given bacterial genomes. Validation and accuracy of the machine learning model was tested using a desiccation assay carried out over three months. To build the model, a literature review was conducted to find genes that were upregulated greater than two-fold during desiccation stress in bacteria. From the review, 2609 genes from 11 papers were found and condensed to 1082 non-homologous and non near-zero variance genes. A second literature search was conducted to identify bacterial species with a known desiccation response, either tolerant or sensitive, and a publicly available genome. Thirty-five desiccation tolerant and 33 desiccation sensitive genomes were chosen and then queried for the previously curated desiccation upregulated genes list. Approximately 176,800 genes were analyzed, and genes with non-zero variance were removed. The remaining 75,982 genes are included in the model (Rogozin et al., 2002). A random forest supervised machine learning approach was used to create a preliminary model for desiccation resistance. The genomes were split into 80% training data and 20% test data and the model was run 100 times with different seeds, 10-fold cross validation, and three repeats. The average accuracy for the 100 iterations of the model was 0.898 ± 0.0266, indicating the

model could accurately predict the desiccation phenotype of the testing data 89.8% of the time. The experimental validation of the desiccation model looked at the viability of 28 bacteria, seven with documented desiccation phenotypes and 21 bacteria with no known desiccation phenotype. For all organisms tested the model had an accuracy of 0.75 demonstrating good model performance.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Defining desiccation

The earth contains many different climates that have organisms evolved to survive in those specific conditions. However, all climates require one molecule for life, water. Some animals and plants can survive in drought or desert conditions, humans, camels, and cacti for example, but mammals cannot survive desiccation. Desiccation tolerance is the ability for organisms to survive at low water activity and then regain function after rehydration. These xerotolerant organisms can survive xeric stress, but do not required it as defined by Lebre et al. Two qualifications must be made to this definition. One, desiccation tolerance is not drought tolerance, a drought is when there is low availability of water in the environment of an organism, while desiccation is low water content inside an organism. Two, there technically is a wide range in desiccation tolerance in prokaryotes ranging from minutes to thousands of years (Potts, 1994). The length of time of desiccation matters, while strictly speaking, some cells can survive being fully desiccated for short periods (e.g., less than a week) and can technically be considered desiccation tolerant. However, such bacteria are not desiccation tolerant for any practical applications. For practical applications and this research, a definition of desiccation tolerance is cells surviving desiccation for longer than three months. The three-month time period relates to the length of time of a standard drought, 3-6 months, this cutoff is explained in greater detail later in this work (Hao et al., 2018).

Bacteria use a variety of strategies to mitigate the effects of desiccation. Individual strains will use multiple redundant systems and different species of bacteria will use different strategies as well. Some adaptations are only in specific organisms and not all organisms. Additionally, some adaptations may be necessary to certain bacteria for survival during desiccation, but not sufficient on their own. Structural, physiological and molecular adaptations have been made by bacteria to help survive desiccation stress. Dormancy and sporulation are structural adaptations in response to many environmental stresses. During desiccation bacteria enter a state of reversible metabolic dormancy wherein they can no longer replicate (Lebre et al., 2017). Spore (or spore-like structure) formation is also an adaptation to protect cells from environmental stresses such as desiccation. Different clades of bacteria have different spore or spore like structure formation that creates a protective shell around the cells, including spores, akinetes, cysts, and myxospores (Laskowska et al., 2020; Rodriguez-Salazar et al., 2017; Reichenbach et al., 1992; Kaplan-Levy et al., 2010). Additional adaptations by some bacteria are the production of exopolysaccharides (EPS) and biofilm formation to hold water and decrease water lost by the cells (Lebre et al., 2017).

Physiological adaptations used during desiccation include cell membrane adaptations, wherein there is an increase in fatty acids in the membrane that become tightly packed to preserve the membrane in a liquid crystalline phase (Lebre et al., 2017). Accumulation (via production or uptake) of compatible solutes, such as trehalose, sucrose, and glycine betaine is also common to stop

the disruption of the membrane during desiccation and result in vitrification (Laskowska et al., 2020). The environmental milleu surrounding the cell can affect desiccation tolerance, for example extracellular sugars, lipids, and proteins can increase desiccation tolerance (Ballom et al., 2020). Metabolic adaptations such as the downregulation of flagellar motility and photosynthesis are induced to reduce the energy used and the reactive oxygen species produced (Lebre et al., 2017). Molecular adaptations such as the presence of late embryogenesis abundant (LEA) proteins in the cells occur in desiccation adapted bacteria, and production of shock-response proteins, ROS scavenger proteins, transcriptional regulators, and sometimes virulence factors occur during desiccation (Lebre et al., 2017).

Desiccation tolerance is an active area of research. LEA proteins have only recently been identified in bacteria after their initial discovery in plants. Research now demonstrates LEA proteins are widely found in different organisms including bacteria, plants, yeasts, and vertebrates. The LEA proteins help protect against protein aggregation and are associated with increased tolerance to environmental stresses such as desiccation and cold stress (Dai et al, 2020). Additionally, one mystery in terms of desiccation research are the organisms in the genus *Deinococcus*. *Deinococcus* is an incredibly robust genus of bacteria that can survive extreme environmental stress including long-term desiccation and high levels of radiation, but the exact mechanisms of this tolerance are yet unknown. In *Deinococcus radiodurans*, LEA protein homologs and regions in proteins with LEA protein homology have been discovered that appear to

improve desiccation tolerance; however, many of the proteins containing LEA regions have unknown functions (Kriško et al., 2010). Recently the term “desiccome” was coined to define the set of proteins that are affected: up and down regulated during desiccation (Ghedira et al., 2018; Potts et al., 2005). Research into desiccation is still ongoing and questions persist as to the mechanisms and genes undergirding desiccation tolerance and the regulation of said mechanisms and genes.

1.2 Relevance of desiccation tolerance

Bacteria are used for many biotechnological applications including agriculture, bioremediation, and industrial uses (Zhanget al., 2020; Clarkeet al., 2020; Yadav et al., 2020; Prasad et al., 2019; Chen et al., 2018). Desiccation tolerance is critical for ease of handling and long-term storage of bacteria. Cells can be shipped around the world, but if they cannot be revived, they cannot perform the desired function. Desiccation tolerance is of special concern for environmentally applied bacteria such as those used for agricultural and bioremediation application. As climate change induced droughts are becoming more common and severe, understanding and enhancing drought and desiccation resistance in cells is necessary (Cook et al., 2014; Dai et al., 2011; Dai et al., 2013).

1.3 Determination of desiccation tolerance

Desiccation response has been widely studied (Potts, 1994; Lebre et al., 2017; Alpert, 2005). Common in the literature are basic studies of the desiccation tolerance of individual bacterial species of interest to a particular research lab. These studies cover a wide range of bacteria; however, they are hard to compare

due to the wide variability in desiccation conditions, timescale, and reported statistics.

Since the advent of transcriptomics, multiple papers report on the regulation of genes during desiccation stress (Ghedira et al., 2018; Cytryn et al, 2007; Gruzdev et al., 2012). The studies include a broad range of bacterial species and a diverse array of genes. In *Bradyrhizobium diazoefficiens*, an agriculturally relevant bacterial species, the desiccome shows genes upregulated for the synthesis of trehalose (*otsA*, *otsB*, *treS*) as well as the upregulation of transcriptional regulators and genes encoding isocitrate lyase, oxidative stress responses, the transport and synthesis of exopolysaccharides, heat shock response proteins, nucleic acid repair and modification enzymes, and genes associated with pili and flagella synthesis (Cytryn et al., 2007). The desiccome of *Frankia alni*, an alder symbiont, identified several changes such as enzymes associated with the cell membrane including mechano-sensitive ion channels and ABC transporters, and production of specific Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)-associated components to alter DNA (Ghedira et al., 2018). In *Salmonella enterica* the desiccome of dormant cells showed few transcripts of heat and cold shock response proteins (Deng et al., 2012). In *Salmonella enterica* serovar Typhimurium the desiccome included most abundantly ribosomal structural genes, and genes involved in amino acid metabolism, energy production, ion transport, transcription, and stress response (Gruzdev et al., 2012). In *Pseudomonas putida* KT2440 the desiccome includes alginate genes, DNA replication and repair genes, trehalose

synthase, and fatty acid and phospholipid metabolism genes (Gulez et al., 2012). In the cyanobacterium *Anabaena* PCC7120, the desiccome includes genes for osmoprotectant metabolism, potassium transport, low temperature stress, and heat shock proteins (Katoh et al., 2004). In *Listeria monocytogenes*, a foodborne pathogen, the desiccome includes energy and oxidative stress genes (ex. *qox*ABCD, *pdh*ABC, *mnt*ABCH), oxidative stress response genes (ex. *Sod*, *kat*, *tpx*, *trx*AB, lmo2390, lmo2830), osmotic stress-related genes (ex. *gbu*ABC, *clp*C, *csp*A, *gro*E), an alternative transcription factor, and long antisense transcripts (Kragh et al., 2019). In *Rhodococcus jostii* RHA1, a soil bacterium the desiccome includes an oxidative stress gene, *dps*1, two genes for sigma factors, SigF1 and SigF3, and the biosynthetic pathway for ectoine (LeBlanc et al., 2008). In *Salmonella* Tennessee and Typhimurium LT2 the desiccome includes fatty acid metabolism genes, stress response, envelope modification genes, and trehalose biosynthesis (Li). In *Chronobacter sakazakii*, a neonatal pathogen, the desiccome includes the trehalose biosynthetic pathway (*otsA* and *otsB*) (Srikumar et al., 2019). Single celled eukaryotes have also been found to contain genes related to desiccation tolerance. In *Saccharomyces cerevisiae*, a yeast, the desiccome includes genes related to fatty acid oxidation and the glyoxylate cycle (Singh et al., 2005). These transcriptomics studies show the surprisingly wide range of genes that are upregulated during desiccation stress in bacteria and yeast. Similar genes may be found in other species not yet studied and the broad range of desiccomes could be modeled with machine learning and used as predictors of desiccation tolerance.

Metagenomics approaches to desiccation are also used to identify mechanisms of desiccation tolerance. In situ analysis of desert bacterial communities with comparative metagenomics analysis identified differences between microbial communities in hot and cold hyper-arid deserts (Le et al., 2016). The metagenome of the hot desert showed more sequences related to metabolism and carbohydrate transport whereas the metagenome of the cold desert showed more sequences for ribosomal structure, replication, translation, repair, sigma factors, and biogenesis (Le et al., 2016). These metagenome differences could indicate a difference in desiccation tolerance strategies.

Beyond desiccation transcriptome research, basic lab testing of the desiccation tolerance of specific strains was conducted in conjunction with transposon mutagenesis (Mandal et al., 2017; Hingston et al., 2017; Humann et al., 2009). Research performed using transposon mutagenesis is an option for determining genes associated with desiccation tolerance; however, there are several issues with this approach. There are typically fewer desiccation tolerant genes discovered than during transcriptomic approaches (Hingston et al., 2017; Humann et al., 2009). Because desiccation tolerance is such a complex phenomenon that has many variations across bacteria, transposon mutagenesis needs to be done to multiple bacterial species that have different mechanisms to be able to define desiccation tolerance genes broadly. Additionally, the experimental setting for the bacterial desiccation can affect which genes appear to contribute to desiccation tolerance (Mandal et al., 2017).

1.4 Addressing the issue of desiccation tolerance

Many research publications and reviews have described individual mechanisms and genes associated with desiccation tolerance in bacteria, but there have been no attempts at creating an overarching framework for the description and prediction of desiccation tolerance in bacteria (Laskowska et al., 2020). Based on this previous research, this study uses machine learning to create a model that predicts desiccation tolerance using bacterial genomes from organisms identified as desiccation tolerant. The prediction accuracy of the model output is then tested using a desiccation assay.

The goal for this research was to standardize, predict computationally, and create an experimental test for desiccation tolerance. By using machine learning modeling, scientists can bypass the complications of determining the exact methods for desiccation tolerance in bacteria and which gene combinations are associated with desiccation tolerance and proceed straight to applying the knowledge of desiccation associated genes to predicting desiccation tolerance in new bacterial species. Previous machine learning research characterized proteins, genes, and organisms computationally with respect to response to a specific stress or metabolic activity (Weimann et al., 2016; Jabłońska et al., 2019; Moradigarav et al., 2018). However, this method has not been applied to desiccation tolerance.

One of the fundamental issues causing a disconnect between science and biology is the human need to categorize and classify everything in a dichotomous or discrete manner. Unfortunately, for many biological categories, especially for

cell properties there are not solid boundaries, very few bacterial properties are all or nothing and much nuance is lost during rigid classification. There are many bacterial responses to stresses that are generally referred to as binary or having a few categories even though true responses to those stresses are on a gradient and not binary, such as salt tolerance, temperature tolerance, antibiotic resistance, oxygen tolerance, and radiation resistance (Ma et al., 2010; Georlette et al., 2003; Jabłońska et al., 2019; Su et al., 2019; Shukla et al., 2007). However, there is a utility (and much history) in categorizing cellular properties, as long as the continuous natures are acknowledged.

There are several classification systems such as Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) in use now for determining the high-level functions and usage of proteins in cells (Ashburner et al., 2000; Kanehisa et al, 2000). Databases have been constructed for the catalog of bacterial species based on specific biological properties such as ionizing radiation resistance (Ryabova et al., 2020). The next step after categorization is computational prediction based on collected data, and that too has started to be explored in the biological property, protein, and gene space. There are several instances of the computational prediction of biological properties recently. In 2016, Weimann and colleagues created the Traitar model (2016). Traitar, “the microbial trait analyzer”, is a fully automated software package for determining phenotypes from a genome sequence. It can predict phenotype classification for 67 traits for the use of numerous substrates as carbon and energy sources, oxygen requirement, morphology, antibiotic susceptibility, proteolysis, and

enzymatic activities. A binary machine learning model was created that can classify bacteria as anaerobic or aerobic based on their genomes (Jabłońska et al., 2019). Antibiotic resistance has been modeled using machine learning several times in multiple bacteria (Moradigarav et al., 2018; Su et al., 2019). Additionally, enzyme substrate specificity has been reliably modeled using machine learning (Robinson et al., 2020). This group used machine learning to predict the substrate specificity of thiolase from structural and physicochemical features. Using a random forest classifier for enzyme activity they obtained an area under the receiver operating characteristic curve of 0.89, which shows excellent prediction accuracy. These studies demonstrate that the classification and computational prediction of biological properties has been successfully carried out and is a viable avenue for the prediction of a more biologically complex property such as desiccation tolerance.

1.5 Machine learning modeling

Due to the amount and complexity of the data collected via transcriptomics, supervised machine learning modeling was chosen to quickly identify patterns in gene abundance across desiccation tolerant and sensitive genomes. Supervised learning is used for classification tasks, and it means that the machine learning algorithm is given both training data and the desired output or classification of the sample in

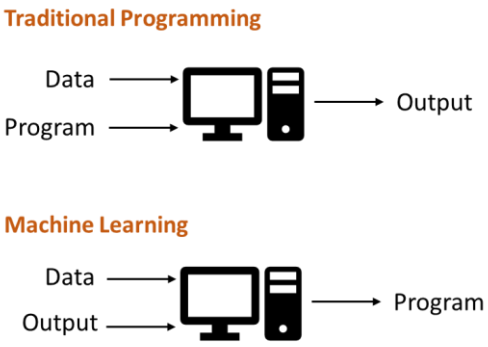

Figure 1: Difference between traditional programming and machine learning. Adapted from Robinson, 2020.

the dataset. Machine learning is different from traditional programming solutions to data analysis (Figure 1). With traditional programming there is a dataset to be analyzed, a program created by the user to analyze the data, and a program that analyzes these features and determines the type result from these properties. When the data is analyzed by the program, a data analysis output is created. For example, someone has a dataset of unknown fruit samples containing information about various features of the fruit, for example, the weight, color, shape, amount of seeds, etc. and a program that identifies fruits based on these features. When the program is run it creates a list of the type of fruit for each sample. This is different from machine learning wherein there is initially a dataset and the desired output from the dataset. This information is analyzed by a machine learning algorithm and a program/model is created that can be used to identify an unknown sample based on their features. To continue the previous example, someone has a dataset of known fruit samples and information about the various features of the fruits used in the example, and when run through the machine learning algorithm a model is created that can identify unknown fruit samples based on their properties.

The machine learning algorithm chosen for this research is the random forest algorithm because of the relatively noisy dataset that was created and the relatively small size of the dataset. The random forest algorithm has less of a chance of overfitting the data than a single decision tree. Machine learning algorithms like neural networks work better with hundreds or thousands of samples and hundreds of thousands of datapoints. Random forest is a machine

learning algorithm used to solve classification problems. For data provided the model created a classification of bacterial species as desiccation tolerant or sensitive on a continuous scale. Random forest then uses ensemble learning, where many classifiers are combined to give the answer to complex problems. In random forest many random subsets of samples and predictors are taken, and a ‘forest” of decision trees is created for each subset. Each decision tree individually is a weak learner, but together the trees of the model “vote” on the classification for each sample via the aggregation of the bootstrap values of the decision trees and the majority rules. In this way the inaccuracies of each tree are minimized, and a stronger model has been trained.

Validation and accuracy of the machine learning model was then tested using a desiccation assay carried out over three months. As there are no standardized methods for testing desiccation tolerance, we created our desiccation apparatus and protocol based on previous descriptions (Farrow et al., 2018; Gruzdev et al., 2011; Vriezen et al., 2006).

1.6 Significance

Better understanding and ability to predict desiccation tolerance are important for culture collections, and industries that culture, store, and ship microbes. As climate change increases severe weather patterns including drought and the necessity of technological intervention, biotechnology is becoming an increasingly widespread solution to a multitude of issues. The model created from this work can be used in two ways. First, as the scope and usage of bacteria in biotechnology increases, this model can be used to predict new

bacterial hosts that can be desiccated and stored long term. Second, it can be used to identify desiccation tolerance in bacteria that can perform long term functions of interest, such as bioremediation. Thus, more competitive, naturally desiccation tolerant species with minimal genetic modification can be used to perform the long-term functions required regardless of periodic droughts.

Chapter 2: Materials and Methods

2.1 Chemicals and Reagents

Potassium acetate (Sigma-Aldrich) was used in the relative humidity (RH) chambers to maintain a constant RH. The following chemicals were used to make media: Calcium chloride dihydrate (Sigma Aldrich), Potassium phosphate dibasic (Sigma Aldrich), LB Broth (BD DifcoTM), R2A Agar (Difco), Casamino acids (Difco), Brain Heart Infusion (BD BBL), D-glucose anhydrous (Mallinckrodt AR), Potassium chloride (Mallinckrodt AR), Nutrient Broth (Oxoid), Marine Broth (Zobell), Tryptone (Research Products International), Yeast Extract (Fermtech), magnesium sulfate 7-hydrate (Bakers Analyzed), Sodium pyruvate (Alfa aesar), Peptone (Fluka), Ammonium sulfate (Fisher scientific), Methanol (Southern Labware).

2.2 Literature curation

2.2.1 Collation of genes upregulated during desiccation stress

A literature review was conducted to find genes associated with desiccation to act as features when building the model. The literature was searched for publications containing transcriptome data that identified genes upregulated in bacteria or yeast during desiccation stress using relevant search terms [“desiccation”, “transcriptome”, “bacteria”, “microarray”, "transcriptomics of desiccation", "transcriptomics of bacterial desiccation", "desiccation bacteria transcript", "desiccation rhizobium transcript", "bacteria sensitive to desiccation"]. Data was also found for genes downregulated during desiccation, but to simplify the model, only upregulated genes were included. Twelve bacterial strains and two yeast strains across 11 publications were found, and 2609 genes identified that were upregulated more than 2-fold in response to desiccation stress. The

gene sequences were downloaded from NCBI and dereplicated using a 50% homology cutoff with the CD-HIT function (Fu et al., 2012). These genes became the list of features in the model.

2.2.2 Assembling bacteria with known desiccation phenotypes

Subsequent to the literature review was identification and classification of bacteria with a publicly available genome plus a known desiccation phenotype, desiccation tolerant or desiccation sensitive. Due to the high variability in methods and results reported in desiccation literature, insufficient data was found to classify bacteria on a continuous scale, thus bacteria were classified based on a binary scale, either desiccation tolerant or desiccation sensitive. While some nuance is lost classifying on a binary scale, Bacteria were classified as desiccation tolerant if they were revived after greater than 90 days of desiccation, were cultured from a desert, or widely recognized as desiccation tolerant in the literature. Species were classified as desiccation sensitive if they could not be revived after 90 days of desiccation or are widely accepted as desiccation sensitive in the literature. Bacteria were removed if the only data reported was the percent of viability after less than 30 days. If the bacteria used in the publication had a publicly available genome, that genome was used, but if the bacterial strain was not specified (common in older publications), a genome for that species was randomly chosen from the NCBI database. The flowchart depicted in figure 1 details the classification determination. These bacteria became the samples in the model. Thirty-five desiccation tolerant and 33 desiccation sensitive strains were identified from the literature.

2.3 Machine learning

2.3.1 Combining the datasets to create matrices

A custom BLAST database was created for the proteome of each bacterium and queried against the list of desiccation relevant genes (Madden, 2002). Only genes with a bit score above 100 and query coverage above 70% were kept. Two matrices were then created. The first was a presence/absence matrix of features (genes) vs samples (bacterial strains) wherein each box had a 1 if the bacteria contained any copies of that gene or a 0 if it contained no copies of that gene. The second was a gene count matrix of features (genes) vs samples (bacteria strains) wherein each box had a count of the number of times that gene appeared in the genome. Genes were removed from the matrices if they had near zero variance using the *nearZeroVar* function from the *caret* package in R (Kuhn, 2008).

2.3.2 Random Forest algorithm

A random forest algorithm was used to create models for both matrices. The data were randomly split 100 times in 80% training and 20% testing sets. R version 3.6.3 and *caret* were used in the evaluation of all models. 10-fold cross validation repeated in triplicate was used to tune model hyperparameters. For the random forest algorithm, model hyperparameters that were tuned included the number of variables randomly sampled as candidates at each split, and the methods for the splitting rules. Relative feature importance was determined using the *varImp* function from the caret package in R (Kuhn, 2008). A second set of random

forest models was created using the top 30 genes of highest variable importance as the features and the same methods detailed above.

2.3.3 Determining and analyzing variable importance plots for the models

For each of the top 30 genes, the average was calculated of the # of instances of that specific gene across all the tolerant species and all the sensitive species (For example AGE85424.1, this gene shows up an average of 3.9 times per genome in the tolerant species and 1.7 times per genome in the sensitive species), and then the ratio was calculated for the tolerant average: sensitive average (3.9:1.7 is a ratio of 2.3:1), therefore AGE85424.1 is 2.3 times more common in the desiccation tolerant species than the desiccation sensitive species).

2.4 Experimental validation of the desiccation model

2.4.1 Bacterial strain determination

The full gene counts model was used to predict desiccation tolerance for 50 strains of bacteria present int the lab of Dr. Wackett (University of Minnesota, MN), 173 strains from DSMZ (Leibniz Institute, Germany), and 1 strain from the Bacillus genetic stock center (Columbus, OH). From those predictions, 28 strains were selected to be used in a 3-month desiccation assay to experimentally validate the model. Strains were chosen that represented a wide taxonomic diversity, were primarily aerobic, non-spore forming, mesophilic, and biosafety level 1. The strains chosen are listed in table A6. The rules as follows were used for the determination of the bacterial strains in the experiment. 1) Bacteria chosen were aerobic as they were to be stored under atmospheric oxygen

content. One exception to this was a *Streptococcus mutans* UA159 that was included in the experiment. The reason for the exception is that *S. mutans* is a microaerophile and it was of interest to test if a microaerophile could survive a high oxygen environment under atmospheric conditions while desiccated as they do contain the reactive oxygen species scavenger superoxide dismutase which could possibly protect them during desiccation (Martin et al., 1984). 2) All the bacteria chosen for the experimental validation were mesophiles because the bacteria were stored at room temperature (22°C) during desiccated-state storage. Most spore forming bacteria were omitted due to difficulty in ensuring sporulation. The exception was two strains of *Bacillus subtilis.* Vegetative *B. subtilis* cells were desiccated to determine if sporulation is the only tolerance mechanism in *B. subtilis* and to test if the predictive model can be “misled” by strains that are genetically engineered. The two sets of *B. subtilis* cells included a wild type *B. subtilis* 168 wherein the cells were grown for one day to promote growth in a symmetric vegetative cycle. Additionally, a non-spore forming *B. subtilis* 168 mutant strain (Bacillus Genome Stock Center, 1S1) was included to ensure the presence of vegetative cells. 3) Bacteria of Biosafety level 1 were chosen to comply with the safety requirements of the lab.

Additionally, several bacteria with reported desiccation tolerance were chosen as a comparison for the desiccation assay used in this research. *Bradyrhizobium diazoefficiens* USDA 110, *Arthrobacter crystallopoietes* ATCC 15481, *Micrococcus luteus* ATCC 4698, and *Deinococcus radiodurans* R1 were chosen as the desiccation tolerant bacteria (Mary et al., 1994; Boylen et al., 1973;

Ujaoney et al., 2017; Mauclaire et al., 2010). *Escherichia coli* DH5alpha, *E. coli* MG1655, and *Shewanella oneidensis* MR-1 were chosen as the desiccation sensitive bacteria (Chen et al., 2018; Daly et al., 2004). Multiple species of *Pseudomonas* were chosen with varying degrees of predicted desiccation resistance were included in the study to determine the sensitivity of the model within a genus.

2.4.2 Desiccation assay setup and quantitation

Each strain was grown to stationary phase in media previously used to grow the bacteria as noted in table A6. Stationary phase bacterial cultures were used as bacteria are able to survive desiccation better in this state (Potts, 1994; Vriezenet al., 2006). Three 1 mL aliquots of each cell culture, samples, were harvested via centrifugation (14,000 g). Each sample was washed twice with equal volumes of sterile deionized water, then resuspended in water. Aliquots (100 µL) of each sample suspension was pipetted into one well in each of five polystyrene 96-well plates (Genesee Scientific). The five 96-well plates were placed in separate desiccation chambers containing a basin of ~200 mL of saturated potassium acetate solution. The saturated potassium acetate solution maintained a RH of 25-35%. The RH was verified using a hygrometer probe (Traceable, VWR). A 3V DC high torque motor (uxcell Store) with a propeller was placed in each desiccation chamber to continuously circulate the air. The dried samples were maintained in a dark cabinet at room temperature and 25-35% RH for 3, 6, 9, or 13 weeks. To determine the initial cell viability each sample was serially diluted in phosphate buffered saline (PBS), samples (10 µL) of each dilution was plated on

their favored media, and colony counts were recorded for each species as they grew, anywhere from one to six days after plating. Determination of the cell viability was conducted at each timepoint, following rehydration (30 min) with PBS (100 µL) and resuspension, and was estimated as described above. Strains were considered no longer viable if fewer than two colonies were counted at the lowest dilution for the three replicates after two successive time points.

2.4.3 Analysis of the desiccation model using the data collected from the experiment

The accuracy, 95% confidence interval, sensitivity, specificity, Area under the receiver operator characteristic curve (AUROC) (calculated using the caret function in R), and F1-score (calculated with the equation F1-score = 2*precision*recall/(precision +recall)) were calculated at each time point (3,6,9,and 13 weeks). The data was analyzed two ways, first with the results of all 28 species tested in the desiccation assay, second with a subset of 24 species. Removed from the second analysis were the *Bacillus subtilis 168* vegetative cells, *B. subtilis* non-spore forming mutant, *Bradyrhizobium japonicum USDA 110*, and *Streptococcus mutans UA159*. The *B. subtilis* strains were removed because they had intentionally been included to test if the predictive model can be “misled” by strains that are genetically engineered or not grown properly. The *B. japonicum* was removed because there is an extensive publication history showing its long-term desiccation survival and its lack of survival in this assay was most likely due to an environmental issue. The *S. mutans UA159* was removed because it is a microaerophile and it could not be

determined if it was killed in the experimental assay via desiccation or the 20% atmospheric oxygen content.

2.5 Creation of the phylogeny tree

A custom BLAST database was created for the genome of each bacterium and queried against the NCBI 16S ribosomal RNA (rRNA) database. The 16S rRNA match for each bacterium with the highest bitscore was then used as the 16S rRNA for each bacteria. The 16s rRNA with the highest bitscore in all cases except one had a percent identity of greater than 97%. The species with a lower percent identity, *Treponema pallidum*, was 89.4% identical with the 16S rRNA for *Treponema denticola*. As the species are in the same genus, and no other species in the model or experiment were in the *Treponema* genus, it was used in the creation of the phylogeny tree. The *read.GenBank* function from the *ape* package was used to download the 16S rRNA sequences from GenBank (Paradis et al., 2019). MegaX was then used to align the sequences via the muscle algorithm with preset settings and create a tree file. The final phylogenetic tree was visualized using the interactive tree of life webpage (Letunic et al., 2021).

2.6 Creation and analysis of the final desiccation model

The newly phenotyped 6 desiccation tolerant and 12 desiccation sensitive species from the desiccation assay were incorporated into the dataset for the desiccation prediction model. Additionally, the *Escherichia coli* MG1655 was converted from desiccation sensitive to desiccation tolerant based on the data

collected in the desiccation assay, for a total of 42 desiccation tolerant species and 44 desiccation sensitive species. A random forest model was created using the same method detailed above for the initial gene counts model creation.

Chapter 3: Results & Discussion

3.1 Machine Learning Workflow

Creation of a machine learning model of desiccation tolerance required the creation of a dataset as the first step. Figure 2 shows the workflow used to create the datasets and use machine learning to model desiccation tolerance. The first step was identifying the genes involved in desiccation tolerance in bacteria. The literature was analyzed for transcriptome studies that reported genes upregulated during desiccation stress. The results of the histogram of desiccation associated genes (Figure 2A) show that the majority of the 2609 genes are upregulated 2- to 20-fold. The gene sequences were downloaded from NCBI and dereplicated using a 50% homology cutoff resulting in 1613 desiccation associated genes. These genes became the features in the model. Additionally, the genes that were upregulated the most were not necessarily the genes of highest variable importance to the model. The gene upregulated the most times, 187 times, ABG92277.1, a propane monooxygenease hydroxylase large subunit from *Rhodococcus jostii* RHA1 was eliminated from the desiccation associated genes during the removal of genes with near zero variance (Leblanc et al., 2020).

The second step in creating the dataset was assembling a list of bacteria with known genomes and a desiccation phenotype classification. Figure 2B shows a simplified version of the decision tree used to determine a binary phenotype (desiccation tolerant or sensitive) of the bacterial species in the model.

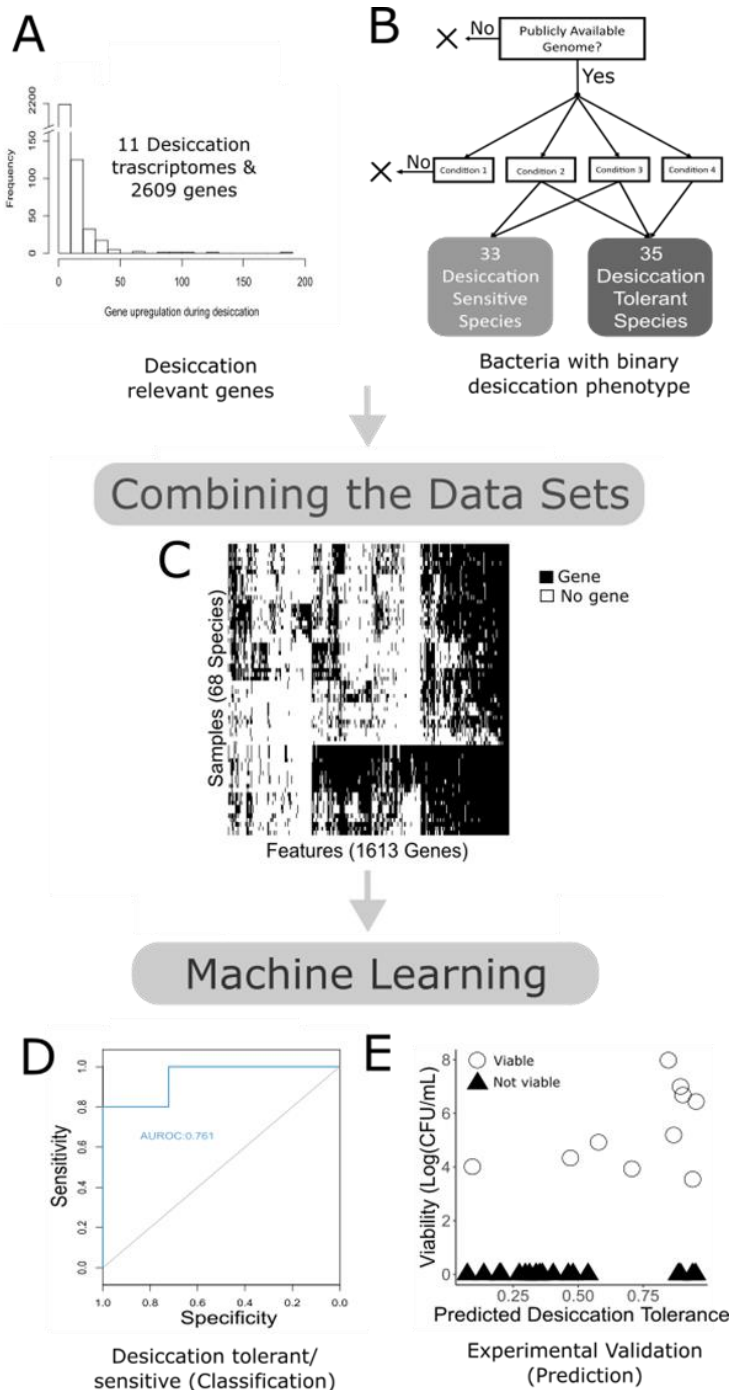

Figure 2: The machine learning workflow. *(A) A histogram of the fold upregulation of the 2609 genes from the literature that were upregulated in response to desiccation stress. (B) A simplified workflow for determining the binary desiccation phenotype (tolerant or sensitive) of bacterial species with published desiccation data. The complete workflow can be seen in figure 3. (C) The binary matrix showing the presence/absence of the 1613 non-homologous genes in the genomes of the 68 bacterial species with a known desiccation phenotype. (D) AUC, Area under the receiver operator curve of the 13-week timepoint of the experimental validation assay. (E) Predicted desiccation tolerance vs Log viability of the 13-week timepoint of the experimental validation assay. The open circles indicate species that maintained viability and the filled triangles indicate species that were no longer viable at 13 weeks. The full dataset can be seen in Figure 6.*

Later, the datasets were combined by querying the genomes against the list of 1613 desiccation associated genes to create matrices of features (1613 genes) vs samples (68 bacteria with known desiccation phenotypes) (Figure 2C). The heatmap in Figure 2C shows a binary matrix of the bacterial species vs the 1613 desiccation associated genes. A random forest algorithm was used to create a classification model, which was analyzed via several metrics including area under the receiver operating characteristic curve (AUROC) (Figure 2D). The model was then used to generate desiccation phenotype predictions for 28 bacterial strains that underwent a desiccation assay to experimentally verify the model. Figure 2E shows the results of the desiccation assay where 8 out of 13 predicted desiccation tolerant strains maintained viability at 3 months and 2 out of 15 predicted desiccation sensitive strains maintained viability at 3 months.

3.2 Genes upregulated during desiccation stress

The model was created based on a feature set of genes (Table 1) identified from previous studies. There were surprisingly few publications that reported transcriptome data during desiccation stress and a complete list of upregulated genes. Genes downregulated during desiccation have also been reported in the literature, and this downregulation is important to maintaining viability of bacteria by reducing unnecessary metabolic activity and regulating other activity (Kragh et al., 2019; Gulez et al., 2012; Ghedira et al., 2018). However, because this research only examines presence or absence of genes, the downregulated genes were not screened. If one bacterium has an ABC transporter for example, Presence or absence of these genes in bacteria shouldn’t indicate desiccation

tolerance or sensitivity, if one desiccation tolerant bacterium has an ABC transporter that is downregulated during desiccation, for example, other bacteria can have that ABC transporter and it does not mean that the transporter is downregulated in response to desiccation stress in those bacteria or that they are more desiccation tolerant. Desiccation resistant organisms with multiple tolerance mechanisms were included; both

Organism	Number of genes	Reference
Bradyrhizobium japonicum USDA 110	693	Cytryn et al., 2007
Listeria monocytogenes 08-5578	430	Kragh et al., 2019
Listeria monocytogenes 568	212	Kragh et al., 2019
Cronobacter sakazakii SP291	365	Srikumar et al., 2019
Rhodococcus jostii RHA1	357	LeBlanc et al., 2008
Pseudomonas putida KT2440	115	Gulez et al., 2012
Saccharomyces cerevisiae BY4743 and commercial strain	104	Singh et al., 2005
Frankia alni ACN14a	81	Ghedira et al., 2018
Salmonella enterica serovar Typhimurium LT2 ATCC 19585	80	Li et al., 2012
Salmonella enterica serovar Tennessee K4643	77	Li et al., 2012
Salmonella enterica serovar Typhimurium SL1344	69	Gruzdev et al., 2012
Anabaena sp. PCC7120	22	Katoh et al., 2004
Salmonella enterica serovar Typhimurium strain ATCC 14028	4	Deng et al., 2012
Total Genes	2609	

Table 1: Number of genes upregulated during desiccation stress that were found in each organism during the literature review.

spore forming and non-spore forming bacteria and a heterocyst former were included. All of the strains with desiccation transcriptome data were different; however, multiple strains of *Salmonella enterica* (Deng et al., 2012; Gruzdev et al., 2012; Li et al., 2012), *Listeria monocytogenes* (Kragh et al., 2019), and *Saccharomyces cerevisiae* (Singh et al., 2005) were included. There was a wide range of conditions for desiccation stress induction, including desiccation at RH from 11%RH to 43%RH, desiccation on different surfaces, desiccation time ranging from 4 hours to 72 hours, one publication desiccated thee bacteria in

peanut oil, and two publications were identified that reported osmotic stress (Deng et al., 2012; Ghedira et al., 2018; Gulez et al., 2012). There is some crossover between genes associated to desiccation stress and genes relevant to osmotic stress. Because the intention was to eventually remove genes with zero or near zero variance, all reported genes were included in earlier searches to be culled from the dataset later. To this end, the transcriptome data from two yeasts, *Saccharomyces cerevisiae* BY4743 and a commercial dry active yeast strain was included in the dataset (Singh et al., 2005). The number of genes from each organism with transcriptome data varied widely from 693 to 4 genes. Upregulation by at least two-fold was chosen to bias towards genes essential to desiccation stress. A total of 2609 genes were pulled from the transcriptome data. After dereplication at a 50% homology cutoff, 1613 genes were used in the model.

The length of time of desiccation is important to note, most of the species in the transcriptome studies were desiccated for only 2-72 hours. Viability is confirmed after desiccation for each of the bacteria in the transcriptome studies, but viability after 72h cannot be extrapolated to predict long-term (2-3 months) desiccation survival without testing. *Listeria monocytogenes* is considered a desiccation tolerant organism but does not fit the definition of desiccation tolerant as defined in this publication, as desiccation tolerance data was only taken up to 48 hours (Kragh et al., 2019). *Pseudomonas putida* KT2440 is considered a desiccation sensitive organism (Manzanera et al., 2020). For genes upregulated during desiccation in a desiccation sensitive organism, it was initially unclear whether

those genes were increasing desiccation tolerance. However, examination of the results of the genes with highest variable importance to the models demonstrate that genes upregulated during desiccation sensitive species can be used to differentiate between desiccation tolerant and sensitive species. Multiple genes from *L. monocytogenes* are present in the top 30 most important genes for both the gene counts and binary models. Specifically, in the gene counts model, the gene identified with the highest score from the variable importance plot is CAC99332.1, an alpha,alpha-phosphotrehalase, a gene that is present three times as much in the desiccation tolerant species as the desiccation sensitive species. This shows that genes from desiccation sensitive species can still be used to model desiccation tolerance.

3.3 Determination of desiccation phenotype in bacteria

Creation of the dataset for the desiccation model required generating a list of bacteria with known genomes and a desiccation phenotype classification (Table 4 & 5). A search of the literature was conducted to identify bacteria with data on the maximum desiccation period for the species. The literature and data discovered in this process was incredibly varied. It was evident that there does not exist a standardized desiccation tolerance assay in bacteria. This is understandable to a certain extent; desiccation response is a complex process in bacteria and the time scale and environmental factors affecting desiccation tolerance are wide-ranging. However, the lack of standardized desiccation data from diverse bacteria potentially limits the predictive power of the model and the machine learning algorithms that can be used. The definition for desiccation

tolerance in the literature are also varied depending on the application and environment of the bacteria in question. Human pathogens that can survive desiccation in a hospital for a month are considered desiccation tolerant due to their potentially large impact on human health, whereas bacteria in the desert that could only survive desiccation for a month would be considered desiccation sensitive due to the extended dry periods encountered in the desert during seasonal change (Hirai et al., 1991). Additionally, environmental conditions such as drying matrix (e.g. trehalose, dextran, blood, feces, milk, soil), drying surface (e.g. steel, glass, plastic, cloth), relative humidity, presence/absence of sunlight, and temperature all can affect the long-term desiccation tolerance of bacteria. Publications in the field of bacterial desiccation span a wide variety of conditions making a standard definition of “true desiccation tolerance” impossible. However, broad categorical binning of desiccation tolerance in bacteria based on time scales relevant to humans is possible.

Originally, the idea was to model desiccation on a scale that would predict days of desiccation for a bacterium with an unknown desiccation tolerance. Initial attempts were made to gather data only on species where the true limit of desiccation tolerance had been discovered (i.e., at one timepoint the bacteria were viable and at the next timepoint they were no longer viable). Unfortunately, there is not sufficient data in the literature to generate a dataset large enough to create such a model that can predict a number of days of desiccation. One of the largest problems is that some desiccation resistance data comparing several species only tested desiccation tolerance for several weeks and reported a the

remaining percentage of bacteria viable after the period of desiccation stress. Such data can only be used to infer relative desiccation tolerance, and no maximum desiccation tolerance can be extrapolated. Another issue to contend with is that some of the studies were published at a time when the specific strain designation was not required to be reported or the species had not yet been identified (Mitscherlich et al., 1984). Different strains of the same species can have different desiccation tolerances, but generally they are similar (Romanovskaya et al., 2002). Thus, for the publications without a strain designation a complete genome from the species was randomly chosen from the NCBI refseq database (NCBI handbook, 2002).

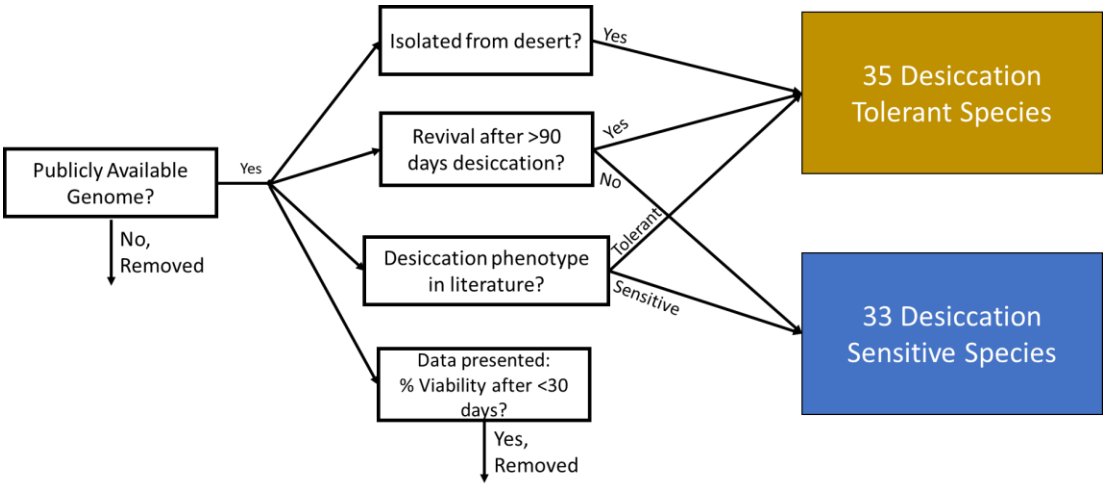

Figure 3: Decision tree for assigning the desiccation phenotype. Only species that have publicly available genomes were included in the study. If the species had been cultivated from a desert environment, they were considered desiccation tolerant. If there was published research showing desiccation testing for more than 90 days and the species was revived it was considered desiccation tolerant. If it was not revived after more than 90 days, it was considered desiccation sensitive. If the species was discussed as desiccation tolerant/sensitive in literature, it was considered tolerant/sensitive. If the only data presented was percent viability of the species after less than 30 days of desiccation, the species was not included in the study.

Due to the lack of data for a ‘days of desiccation’ model a decision tree was created to assign a binary classification for desiccation tolerance of the bacteria (Figure 3). Seasonal droughts typically last three to six months (Hao et al., 2018). In desiccation data published in the literature, there is a natural separation gap in the reported desiccation exposure between about 1-3 months. Therefore, in this research, a cutoff point of 3 months was chosen to delineate a binary classification of desiccation tolerant/sensitive. In data reporting percent viability after fewer than 30 days, the desiccation tolerance could not be extrapolated to 90 days, and the species was removed. Species cultured from a desert were considered desiccation tolerant. Some species are designated as desiccation tolerant or sensitive in the literature but there is not long-term desiccation data. These bacteria were included in the model and were considered the desiccation phenotype designated in the literature. Organisms from the literature that were greater than 50% viable after more than 30 days of desiccation were considered desiccation tolerant. Studies of bacteria dried in environments intended to improve desiccation tolerance (dextran, trehalose, etc.) were excluded from the study because the effects of xeroprotectant chemicals on gene expression is not known. These parameters resulted in inclusion in the analysis of a relatively even split of 35 desiccation tolerant species and 33 desiccation sensitive species (Figure 3).

3.4 Random Forest Machine Learning Algorithm

The two data sets including desiccation related genes and bacteria with known desiccation phenotype were combined into a gene counts matrix and a binary

matrix (Figure 2C). When near zero variance genes were removed, the gene count matrix contained 1087 relevant genes while the binary matrix had 1026 genes. The high number of desiccation associated genes demonstrates the complexity of desiccation tolerance, and a machine learning model was necessary to be able to determine the full contribution of all the desiccation associated genes.

A random forest algorithm was used to predict desiccation tolerance in bacteria using a binary designation. Measurements of model accuracy can be judged on the same scale from 0.5 to 1.0. Accuracy below 0.5 is non-predictive while accuracy of 0.5 is the same accuracy as a coin toss. Accuracy below 0.7 is suboptimal performance, accuracy of 0.70 to 0.80 indicates a good performance of the model, accuracy of greater than 0.8 shows excellent performance by the model, and accuracy of 1.0 means that the model is a perfect classifier. An average testing set classification accuracy of 88.2±9.1% and 87.3±8.5 % was obtained from 100 random training-testing dataset splits for the gene counts model and binary model respectively (Figure A1). This indicates that models using all genes have an excellent classification accuracy. Variable importance plots of the top 30 genes of highest variable importance in the models were determined. To simplify the models, the random forest algorithm was run using the top 30 most important genes as features for each dataset creating two additional models. Because of the random nature of the algorithm, the variable importance plots are not reproducible across different models trained on the

same dataset. The classification accuracies of example analysis models are shown in Figure A1.

For each random training-testing dataset split a different set of top 30 genes of highest variable importance can be created. Example top 30 variable importance plots for the gene counts and binary models are shown in Figure 4. In the remainder of the text, the gene counts model will be called the “full gene counts model” and the binary model will be called the “full binary model.” The average classification accuracy of the training set for the gene counts top 30 gene and binary top 30 genes model increased to 90.9±0.4%, and 91.0±0.3%, respectively, and the test set for the gene counts top 30 gene and binary top 30 genes model increased to 96.6±5.3%, and 96.2±3.1%, respectively (Figure A1B & A1D). The high training and test accuracy could indicate overfitting in the top 30 gene models. Thus, we decided to use the full gene counts model with the full set of genes to estimate and verify the predictive capabilities of our bioinformatics desiccation modeling approach. However, because of the high prediction value of the top 30 genes models the experimental data were also analyzed with these models.

3.5 Results of the variable importance plots

It is necessary to understand how the model identifies variable importance. The model is naïve, it does not know what desiccation is and the genes it identifies may occur at a higher frequency in desiccation tolerant or sensitive species and not be associated with desiccation. For example, some of the secondary metabolic clusters such as those related to polyketide synthesis that are more

common in the desiccation tolerant species might have higher variable importance because more soil organisms are desiccation resistant and also have more biosynthetic gene clusters and larger genomes. The organism of origin for the genes with highest variable importance is interesting as well. Only genes from 6 of the 14 species with transcriptome data appeared in the top 30 genes of variable importance. The reason for this is unknown, but it is possible the genes from those organisms are more widespread through the genomes of the bacteria in the model.

In addition to examining the genes with highest variable importance, the average presence of the genes in the tolerant species and the sensitive species were used to calculate the ratio of presence in tolerant species vs presence in sensitive species to see how much more common the top 30 genes were in the tolerant species as opposed to the sensitive species (Table 6 & 7). For each of the top 30 genes, average occurrence in tolerant vs sensitive species was calculated. Some genes have ratios greater than 1 indicating they are more common in desiccation tolerant species. Some genes have a ratio of less than 1 indicating that gene is more common in desiccation sensitive species than desiccation tolerant species. It was not expected that genes would have ratios <1 as the initial prediction was that desiccation tolerant species would contain more desiccation associated genes than desiccation sensitive species. However, presence of organisms with gene ratios <1 show that desiccation sensitive species have more copies of some desiccation upregulated genes than desiccation tolerant species.

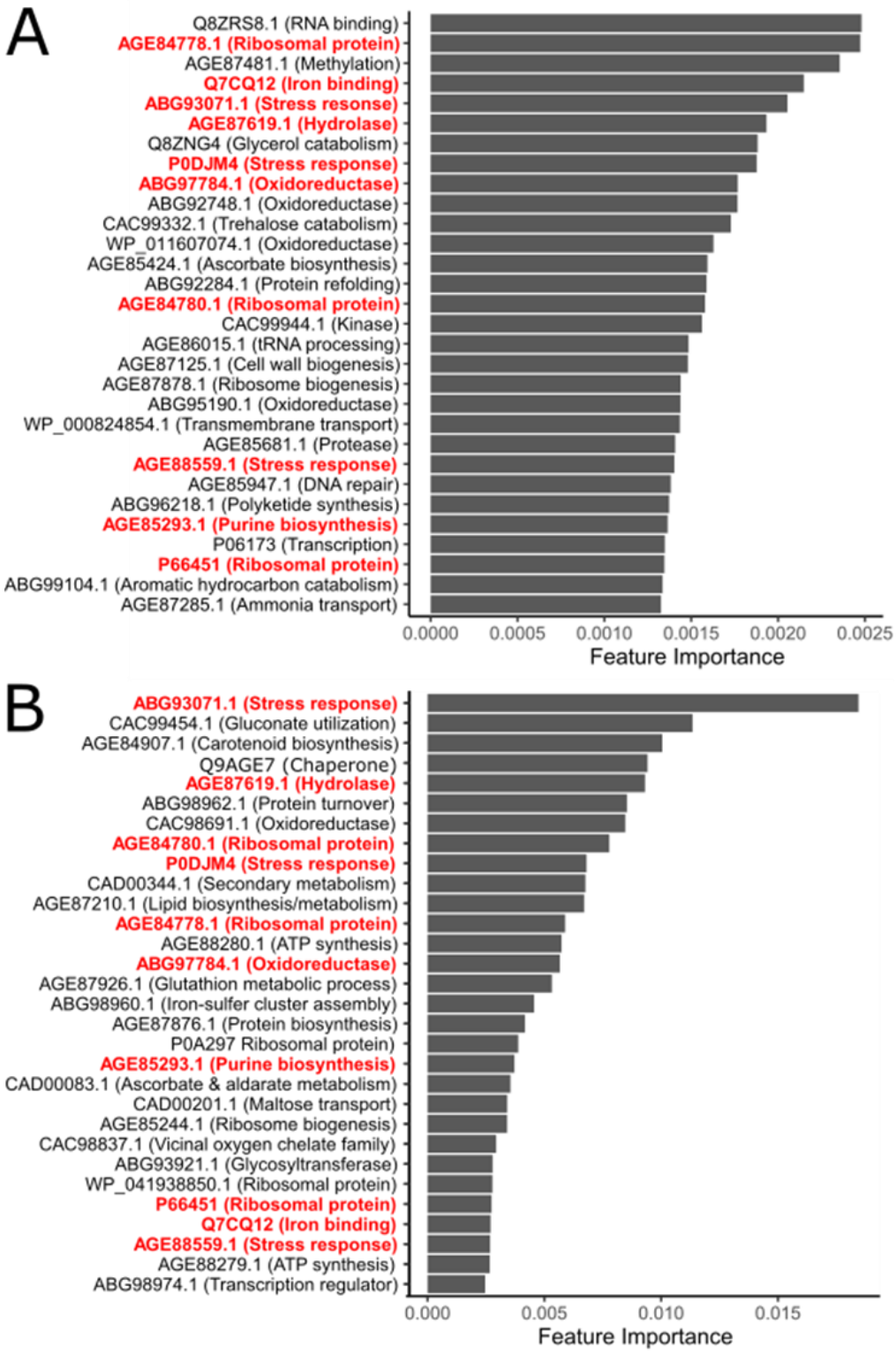

Figure 4: Example variable importance plots *showing the top 30 most important features (genes) in order of importance to (A) the gene counts model and (B) the binary model. The protein accession number is given first followed by its general function. The 10 genes shared by the models are bolded in red.*

Additionally, the genes with the highest variable importance are not necessarily the genes with the highest or lowest ratios. Thus, the genes of highest variable importance are not necessarily the most important to long term desiccation.

However, there are multiple genes that are more common in the desiccation tolerant bacterial species that can be directly linked to desiccation tolerance. The first category of genes with high variable importance in the models are oxidoreductase genes. Oxidoreductase enzymes catalyze the transfer of electrons from one molecule to another. They are common enzymes in cells and are part of pathways some of which are known. Heat shock proteins are also important to the model. Heat shock proteins are known as general stress response genes and for their role in increasing desiccation tolerance (LeBlanc et al., 2008; Garbuz et al., 2017). CAC99332.1, an alpha,alpha-phosphotrehalase codes for an enzyme important in the synthesis of trehalose (Bhumiratana et al., 1974). Trehalose is one of the most important osmolytes that can increase desiccation tolerance in cells (Lebre et al., 2017). Therefore, the presence of a gene in the trehalose catabolism pathway is not unexpected (Lebre et al., 2017).

The model identified several genes that code for secondary metabolites that are important in the models. These include genes relevant to polyketide synthesis, aromatic hydrocarbon metabolism, ascorbate biosynthesis, and carotenoid biosynthesis. As stated earlier, some secondary metabolic clusters such as polyketide synthesis might be more common in the desiccation tolerant species because there were many soil and environmental organisms among the desiccation tolerant species. The soil and environmental organisms evolved in

environments with multiple stresses (including desiccation, radiation, and temperature stresses), are desiccation resistant, have more biosynthetic gene clusters and larger genomes. These secondary metabolites are not necessarily relevant to desiccation tolerance; however, ascorbic acid and beta-carotene are antioxidants which could be upregulated to help protect the bacteria from the DNA damage that occurs during desiccation (Lebre et al., 2017; Kranner et al., 2005).

Additionally, multiple genes associated with iron-sulfur clusters were identified with high variable importance in the full gene counts and full binary models; however, some of these genes are more common in desiccation sensitive species. Most bacteria accumulate more iron than manganese. Iron is incorporated into many enzymes that have iron-sulfur clusters, heme, and mono-nuclear iron centers like naphthalene dioxygenase (Daly et al., 2004). Some organisms use manganese in place of iron for many functions intracellularly (Daly et al., 2004). It has been observed that substitution of manganese for iron correlated very well with radiation resistance (Ghosal et al., 2005). Interestingly, most radiation resistant bacteria are also desiccation tolerant and there is thought to be some cross-over in mechanisms (Ghosal et al., 2005).

3.6 Selection of bacteria for the desiccation assay

3.6.1 Desiccation predictions

To experimentally verify the model, bacteria identified in the model were tested in a desiccation assay to compare the desiccation survival to desiccation tolerance predictions. We conducted a literature search for bacterial species with known genomes and calculated desiccation prediction scores for approximately 250 bacteria using four models. Predictions were binary but were ranked on a scale from 0 to 1. A score of 0.00-0.49 indicates the species was predicted to be desiccation sensitive, and a score of 0.51-1.00 indicates the species was predicted to be desiccation tolerant. The closer the prediction value was to 0 or 1 the more certain the model was in the prediction. A prediction of 0.5 would mean the model could not predict a desiccation phenotype. We then ranked the bacteria

Bacterial Strain	Desiccation Model Prediction
Arthrobacter crystallopoietes DSM 20117	0.95
Rhodococcus B1	0.95
Bacillus subtilis 168 vegetative cells	0.94
Bacillus subtilis 168 1S1 non-spore mutant	0.94
Arthrobacter aurescens TC1	0.90
Bradyrhizobium diazoefficiens USDA 110	0.89
Listeria grayi DSM 20601	0.89
Sinorhizobium medicae WSM 419	0.87
Deinococcus radiodurans R1	0.85
Methylobacterium extorquens AM1	0.85
Micrococcus luteus ATCC 4968	0.71
Sphingomonas yanoikuae b1	0.58
Acidobacterium capsulatum DSM 11244	0.54
Pedobacter heparinus DSM 2366	0.48
Paraburkholderia xenovorans LB400	0.47
Ralstonia eutropha H16	0.46
Herbaspirillum sp. CAH-3	0.40
Streptococcus mutans UA159	0.36
Flavobacterium aquatile F36	0.35
Pseudomonas mendocina	0.34
Comamonas sp. CAH-2	0.31
Marinobacter hydrocarbonoclasticus DSM 8798	0.30
Pseudomonas chlororaphis DSM 50083	0.27
Pseudomonas putida F1	0.20
Alkanivorax borkumensis sk2	0.20
Escherichia coli DH5alpha	0.14
Escherichia coli MG1655	0.09
Shewanella oneidensis MR-1	0.07

Table 2: The bacteria used in experimental validation of the model and their predicted desiccation resistance. *Prediction scores >0.5 indicate the species are desiccation tolerant (green). Scores <0.5 are predicted to indicate desiccation sensitivity. Scores closer to 0.5 indicate the model is less certain in the prediction.*

based on their predictions and selected organisms across the range of predictions. This helped determine if the model works better for determining desiccation sensitive or tolerant species and how accurate the model for predictions around 0.5. Predictions with the full gene count model for the bacteria can be seen in Table 2.

To determine the models' usefulness for diverse bacteria, bacteria were chosen from a wide range of phyla and classes. The similarity in genetic diversity between the species in the model and the experimental species can be seen in Figure A2 and the genetic diversity within the experimental strains can be seen in Figure 5. In Figure A2, the experimental species are distributed fairly evenly throughout the species from the model.

3.6.2 Genetic proximity of the bacteria included in the model and experiment

The bacteria included in the model cover seven phyla and five classes of proteobacteria. There is variability within the larger classes of bacteria, as the classes include bacteria of both desiccation sensitive and tolerant phenotypes. For example, among the alpha-proteobacteria there are six desiccation tolerant bacteria and three desiccation sensitive bacteria, among firmicutes there are seven desiccation tolerant bacteria and two desiccation sensitive bacteria, and among gamma-proteobacteria there are six desiccation tolerant bacteria and 18 desiccation sensitive bacteria. The diversity of organisms included in the model helps to ensure the model predictions are based on the presence and regulation of desiccation genes to eliminate the possibility of grouping species as

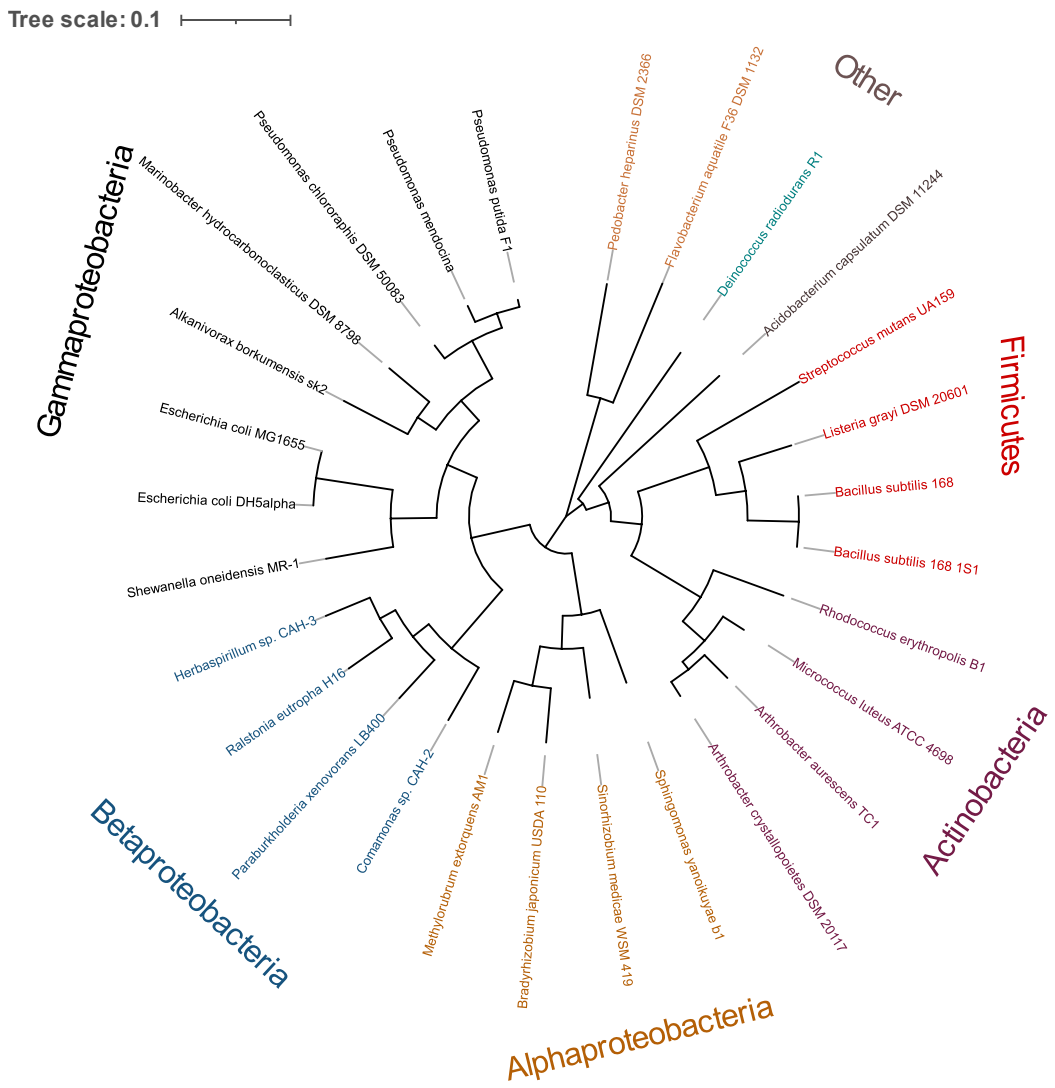

Figure 5: Phylogeny tree for experimental species color-coded by phylum/class*. In the other subset of organisms, Acidobacterium capsulatum DSM11244 is in the phylum Acidobacteria, Deinococcus radiodurans R1 is in the phylum Deinococcus-Thermus, and Pedobacter heparinus DSM 2366 and Flavobacterium aquatile F36 DSM 1132 are in the phylum Bacteroidetes. The tree scale showing genetic distance is shown in the upper left-hand corner.*

desiccation tolerant or sensitive based solely on the phylum. The bacteria included in the experimental validation study cover six phyla, and three classes of proteobacteria (Figure 5). There is one phylum in the experiment that was not included in the model, Acidobacteria to determine how accurate the models

predict desiccation tolerance in bacteria in a phylum they have not yet encountered.

3.7 Results of the model validation experiment

Experimental validation of the desiccation prediction algorithm determined the post-desiccation viability of 28 bacteria, (seven bacteria with previous desiccation data and 21 bacteria with no known desiccation phenotype) (Figure 6). Of the 13 species predicted to be desiccation tolerant by the model, three did not survive three months of desiccation. Of the 15 species predicted to be desiccation sensitive 13 did not survive three months of desiccation (Figure 6 & A3). Four bacteria previously determined to be desiccation tolerant in the literature were included in the experimental validation assay. *Deinococcus radiodurans* R1 known to be the most desiccation tolerant bacteria, maintained its viability over the 13-week assay period. Another species known to be desiccation tolerant, *Micrococcus luteus* ATCC 4698 lost 3-logs of viability gradually during the first six weeks but maintained its viability for the remainder of the assay. *Bradyrhizobium diazoefficiens* was viable at the start of the assay but did not survive three weeks of desiccation. Multiple studies confirm the long-term desiccation tolerance of *Bradyrhizobium diazoefficiens* (Mary et al., 1994; Antheunisse et al.,1981). However, in these reports many different media are used to grow *Bradyrhizobium diazoefficiens*. It seemed to grow well on tryptone-yeast extract media in this assay, but this media may not be ideal for the *B. diazoefficiens* and may have affected its survival during desiccation. The last bacterium considered desiccation tolerant in the literature that was included in the validation,

Arthrobacter crystallopoietes ATCC 15481, survives up to six months when mixed with sand and air dried (Boylen et al.,

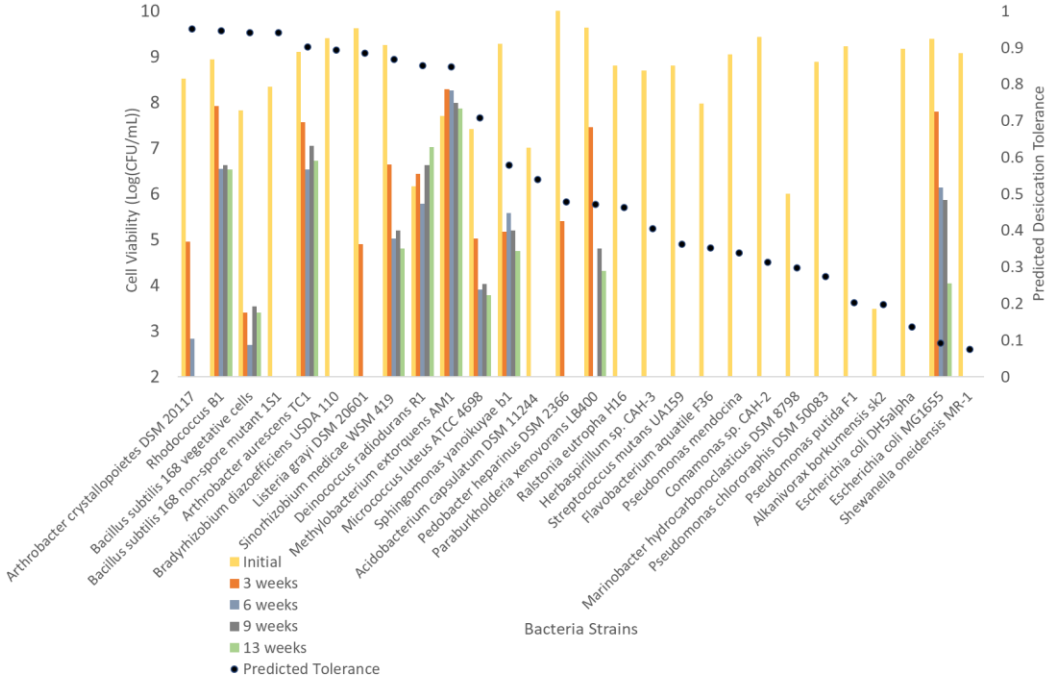

Figure 6: Desiccation assay results. *The graph is organized by predicted desiccation tolerance with the highest on the left and lowest on the right. Of the 13 species predicted desiccation tolerant, three did not survive 13 weeks of desiccation. Of the 15 species predicted desiccation sensitive 13 did not survive 13 weeks of desiccation.*

1973). In this experimental validation it survived six to nine weeks. This could be due to a difference in the drying substrate, in this study it was dried on plastic whereas previously it was dried on sand or glass beads, or a difference in media. The previous study used a glucose mineral salts media, while nutrient broth was used in the current experiment.

The *Listeria greyi* was strongly predicted to be desiccation tolerant, most likely due to the presence of the *L. monocytogenes* (phenotyped as desiccation tolerant in the model) genomes in the model. However, it only survived three to six weeks of desiccation. The reason *Listeria monocytogenes* is labeled

desiccation tolerant in the literature is most likely due to its detection in a food packaging or medically relevant settings, which determine bacteria to be desiccation tolerant after shorter periods of time. However, it is not desiccation tolerant by the definition outlined in this work as it was not viable after three or more months of desiccation. The *Bacillus subtilis* non-spore mutant cells lost viability prior to the three-week time point as expected. Without the spore coat to protect them, the vegetative cells are incredibly sensitive to desiccation stress. The *Bacillus subtilis* 168 WT “vegetative” cells had low cell viability at all follow-up time points after the initial viability. Based on the complete lack of viability of the non-spore forming mutants, it is suspected that some spores developed, resulting in the low levels of viability in the *Bacillus subtilis* WT cells.

There were three negative desiccation sensitive controls. The *E. coli* DH5alpha and *Shewanella oneidensis* MR-1 did not survive to three weeks. The third desiccation sensitive control, *E. coli* MG1655, a WT strain, was shown previously to not be viable after 28 days desiccation (Chen et al., 2018). Here it survived three months with an ~5-log drop. This difference in desiccation tolerance may be due to variation in experimental apparatus and indicates a standardized desiccation assay is necessary. The other desiccation tolerant bacteria that was predicted desiccation sensitive, but near the middle of the desiccation tolerance range (0.47), was *Paraburkholderia xenovorans* LB400. It was recovered at the 13-week time point.

One of the drawbacks of the laboratory testing approach chosen was that during the desiccation assay, there was a fan constantly moving the air around in the

controlled relative humidity boxes during the assay. This may have produced a low but consistent vibration of the cells possibly increasing the damage caused during desiccation and reducing viability. Additionally, the motors started to fail after constantly running for about nine weeks and there was a slight residue in the boxes from the motor, but not in the assay plate, that could have interfered with the assay. Future studies testing desiccation tolerance should either dry the bacteria in a biosafety cabinet with the fan on and then transfer them to boxes with a controlled relative humidity or dry them for three days with the fan/motor setup and then turn the fans off for the remainder of the assay.

There was a general agreement between predicted and experimentally determined desiccation tolerance. Although some of the control species did not result in expected outcomes, this variation could be due to non-ideal growth media for recovery of damaged cells or growth phase of the organism prior to desiccation. Genetic engineering of non-wild type cells may induce errors in the model prediction.

3.8 Accuracy of the model

The results of the model were analyzed for accuracy using the 95% confidence interval, sensitivity, specificity, AUROC, and the F1-score (Table 3). An analysis was done for the results of each timepoint and on two sets of data, all the experimental data, and experimental data with selected non-ideal samples removed. All the measurements of model accuracy can be judged on the same scale from 0.5 to 1.0. Accuracy below 0.5 is non-predictive while accuracy of 0.5 is the same accuracy as a coin toss. Accuracy below 0.7 is suboptimal

performance, accuracy of 0.70 to 0.80 indicates a good performance of the model, accuracy of greater than 0.8 shows excellent performance by the model, and accuracy of 1.0 means that the model is a perfect classifier. Accuracy

	Weeks of Desiccation	Accuracy	95% confidence interval		Sensitivity	Specificity	AUROC	F1-score
All experimental data	3	0.79	0.59	0.92	0.80	0.77	0.79	0.77
	6	0.82	0.63	0.94	0.78	0.90	0.84	0.78
	9	0.75	0.55	0.89	0.72	0.80	0.76	0.70
	13	0.75	0.55	0.89	0.72	0.80	0.76	0.70
Experimental data with selected samples removed	3	0.83	0.63	0.95	0.92	0.75	0.83	0.82
	6	0.88	0.68	0.97	0.87	0.89	0.88	0.84
	9	0.79	0.58	0.93	0.80	0.78	0.79	0.74
	13	0.79	0.58	0.93	0.80	0.78	0.79	0.74

Table 3: Accuracy of the full gene counts model as verified by the experimental data at 3,6,9, & 13 weeks. *The data was analyzed two ways, first with the results of all 28 species tested in the desiccation assay, second with 24 species. Removed from the analysis were the B. subtilis 168 vegetative cells, B. subtilis non-spore forming mutant, B. diazoefficiens USDA 110, and S. mutans UA159. The B. subtilis strains were removed because they had originally been used to test the model. The B. diazoefficiens was removed because there is an extensive publication history showing its long-term desiccation survival and its lack of survival in this assay was probably an environmental issue. The S. mutans UA159 was removed because it is a microaerophile and it is unknown if it was killed by the desiccation or the 20% atmospheric oxygen content.*

measurements of the models were very similar across the various metrics. All the models had an accuracy greater than 0.7 showing good performance by the model. Accuracy was calculated for each timepoint to see which timepoint had the highest accuracy. Accuracy increased from the 3-week timepoint to the 6-week timepoint, but then deceased for the 9- and 13-week timepoints. This indicates that the model works best to determine which species are desiccation tolerant and sensitive up to six weeks of desiccation. Accuracy scores were consistently higher in the dataset with the non-ideal samples removed, with the

highest score being the 6-week timepoint of the dataset with non-ideal species removed.

Sensitivity is the ability of the model to correctly identify true positives, which in this experiment is the desiccation tolerant species. Specificity is a measure of the ability of the model to correctly identify true negatives, the desiccation sensitive species. Sensitivity and specificity are very similar for the 13-week timepoint of the subset of the data, 0.80 and 0.78 respectively, showing the model works equally as well for predicting desiccation tolerant and sensitive species.

Accuracy across the models was similar with the full gene counts model and the top 30 binary model performing the best (Table 10).

While the accuracy is similar for both top models, the full gene counts is preferable to ensure the use of all the genes. If the bacteria are lacking one gene from the top 30 binary genes model it would have a much bigger effect on their prediction score. Use of the full gene counts model gives a more well-rounded score because it is based on more genes.

3.9 Creation of the final model

A final model was made to incorporate all the data collected from the literature review and desiccation assay into the model. All of the data from the desiccation phenotypes species from the literature review and the desiccation assay were used in the training set for a gene counts random forest model to include the most species and improve the applicability of the model. The two desiccation tolerant controls that did not survive in the assay were kept as desiccation

tolerant species, but the *E. coli* MG1655 was switched to desiccation tolerant based on the data from the assay. An average training set classification accuracy of 87.9 ± 0.87% was obtained from 100 random iterations of the random forest algorithm (Figure 7). This model and the code required can be found online at https://github.com/clips002/Desiccation-Modeling and used for desiccation predictions by other researchers with bacterial genomes.

3.10 Conclusions

Desiccation tolerance is a complex process and desiccation research is ongoing. Machine learning models were trained to predict desiccation tolerance in bacteria. A standardized desiccation assay was created and used to experimentally verify the model. There is general agreement between predicted and experimentally-determined desiccation tolerance for the above random forest models with a maximum accuracy of 0.79 for the full gene counts model at 13 weeks.

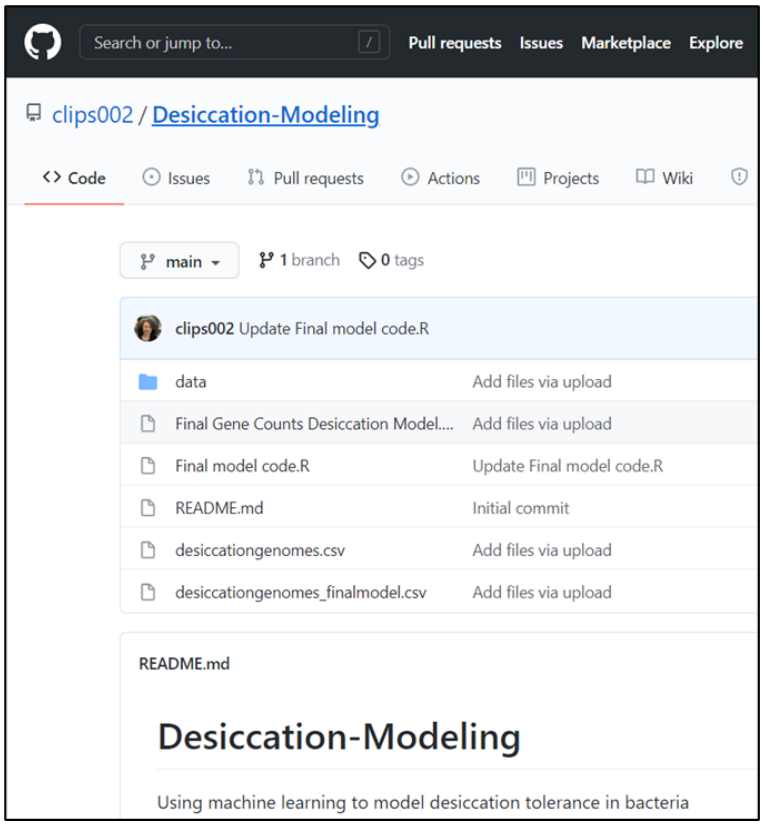

Figure 7: Webpage where the code required to run the model can be found online at https://github.com/clips002/Desiccation-Modeling

In general, Actinobacteria are among the most desiccation tolerant, and gamma-Proteobacteria the most desiccation sensitive bacteria in the model and this was experimentally verified.

Each of the two models gave a ranked set of genes of highest variable importance, wherein there was overlap of ~30% (10/30).

Of the important genes shared by both models or highlighted as the top gene by one, widespread mechanisms of desiccation tolerance related to turning on generalized stress responses and mitigating against free iron release, likely to prevent the formation of damaging oxygen radicals.

This model can be used to predict desiccation tolerance in bacterial species that have not been experimentally tested for desiccation tolerance.

3.11 Next steps

It would be interesting to investigate the ability for the models created in this work to predict tolerance to stresses beyond desiccation. Because there are genes such as heat shock proteins and DNA protection and repair genes that help protect against multiple stresses (desiccation, radiation, temperature, acid, etc.) there is a possibility that the models created herein could be used to predict tolerance for other stresses.

Additionally, alterations could be made to the model to sensitize it further to the phylums of bacteria that have demonstrated variable desiccation tolerance. For some phylums, every species with desiccation data was the same predicted desiccation tolerance. For example, all actinobacteria found were determined to

be desiccation tolerant, and all betaproteobacteria were determined to be desiccation sensitive. A simplified model could be created that only included species from a phylum that contained variable desiccation phenotypes. Then when a bacterium needed a desiccation phenotype prediction the model would look first at the phylum of the bacterium in question and if it was from a phylum where all of the bacteria had the same desiccation phenotype, that phenotype would be assigned to that bacteria. If it was not, a prediction would be run using the model with the variable phenotype bacteria.

Additional experiments could be run to obtain desiccation transcriptome data on a wider variety of bacterial species to increase the pool of desiccation associated genes.

To be able to predict desiccation tolerance in other fields models can be trained for other organisms such as archaea, plants, animals, and fungi. A combined model could also be created to determine shared genes across all types of life that increase desiccation tolerance.

Bibliography

Ahn, S.J., Lemos, J.A., and Burne, R.A. (2005) Role of HtrA in growth and competence of Streptococcus mutans UA159, *J Bacteriol* **187**: 3028-3038.

Alpert, P. (2005) The limits and frontiers of desiccation-tolerant life, *Integr Comp Biol* **45**: 685-695.

Antheunisse, J., de Bruin-Tol, J.W., and van der Pol-van Soest, M.E. (1981) Survival of microorganisms after drying and storage, *Antonie Van Leeuwenhoek* **47**: 539-545.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, *Nat Genet* **25**: 25-29.

ATCC (2021) *Arthrobacter aurescens* Phillips 13344.

ATCC (2021) *Arthrobacter crystallopoietes* Ensign and Rittenberg 15481 ™.

ATCC (2021) *Cupriavidus necator* 17699 ™.

ATCC (2021) *Marinobacter hydrocarbonoclasticus* Gauthier et al. 49840.

ATCC (2021) *Micrococcus luteus* (Schroeter) Cohn 4698.

ATCC (2021) *Pseudomonas chlororaphis* subsp. *chlororaphis* (Guignard and Sauvageau) Bergey et al.9446

ATCC (2021) *Pseudomonas putida* (Trevisan) Migula 700007.

ATCC (2021) *Shewanella oneidensis* Venkateswaran et al.700550.

Bacdive (2021) Bacdive 3852.

Baev, M.V., Baev, D., Radek, A.J., and Campbell, J.W. (2006) Growth of Escherichia coli MG1655 on LB medium: determining metabolic strategy with transcriptional microarrays, *Appl Microbiol Biotechnol* **71**: 323-328.

Bailly, X., Giuntini, E., Sexton, M.C., Lower, R.P., Harrison, P.W., Kumar, N., and Young, J.P. (2011) Population genomics of Sinorhizobium medicae based on low-coverage sequencing of sympatric isolates, *ISME J* **5**: 1722-1734.

Ballom, K.F., Tsai, H.C., Taylor, M., Tang, J., and Zhu, M.J. (2020) Stability of Listeria monocytogenes in non-fat dry milk powder during isothermal treatment and storage, Food Microbiol 87: 103376.

Bhumiratana, A., Anderson, R.L., and Costilow, R.N. (1974) Trehalose metabolism by Bacillus popilliae, J Bacteriol 119: 484-493.

Boylen, C.W. (1973) Survival of Arthrobacter crystallopoietes during prolonged periods of extreme desiccation, *Journal of bacteriology* **113**: 33-37.

Chaibenjawong, P., and Foster, S.J. (2011) Desiccation tolerance in Staphylococcus aureus, *Arch Microbiol* **193**: 125-135.

Chater, K.F. (2001) Regulation of sporulation in Streptomyces coelicolor A3(2): a checkpoint multiplex?, *Current Opinion in Microbiology* **4**: 667-673.

Chen, A.I., and Goulian, M. (2018) A network of regulators promotes dehydration tolerance in Escherichia coli, *Environ Microbiol* **20**: 1283-1295.

Chen, G.Q., and Jiang, X.R. (2018) Next generation industrial biotechnology based on extremophilic bacteria, *Curr Opin Biotechnol* **50**: 94-100.

Clarke, L., and Kitney, R. (2020) Developing synthetic biology for industrial biotechnology applications, *Biochem Soc Trans* **48**: 113-122.

Cook, B.I., Smerdon, J.E., Seager, R., and Coats, S. (2014) Global warming and 21st century drying, *Climate Dynamics* **43**: 2607-2627.

Cytryn, E.J., Sangurdekar, D.P., Streeter, J.G., Franck, W.L., Chang, W.S., Stacey, G., et al. (2007) Transcriptional and physiological responses of Bradyrhizobium japonicum to desiccation-induced stress, *J Bacteriol* **189**: 6751-6762.

Dai, A. (2011) Drought under global warming: a review, *WIREs Climate Change* **2**: 45-65.

Dai, A. (2013) Increasing drought under global warming in observations and models, *Nature Climate Change* **3**: 52-58.

Dai, J., Gao, K., Yao, T., Lu, H., Zhou, C., Guo, M., et al. (2020) Late embryogenesis abundant group3 protein (DrLEA3) is involved in antioxidation in the extremophilic bacterium Deinococcus radiodurans, *Microbiol Res* **240**: 126559.

Daly, M.J., Gaidamakova, E.K., Matrosova, V.Y., Vasilenko, A., Zhai, M., Venkateswaran, A., et al. (2004) Accumulation of Mn(II) in Deinococcus radiodurans facilitates gamma-radiation resistance, *Science* **306**: 1025-1028.

Deng, X., Li, Z., and Zhang, W. (2012) Transcriptome sequencing of Salmonella enterica serovar Enteritidis under desiccation and starvation stress in peanut oil, *Food Microbiol* **30**: 311-315.

Desikan, K.V., and Sreevatsa (1995) Extended studies on the viability of Mycobacterium leprae outside the human body, *Lepr Rev* **66**: 287-295.

DSMZ (2021) Acidobacterium capsulatum DSM 11244.

DSMZ (2021) Alcanivorax borkumensis DSM 11573.

DSMZ (2021) Bacillus subtilis DSM 23778.

DSMZ (2021) Flavobacterium aquatile DSM 1132.

DSMZ (2021) Listeria grayi DSM 20601.

DSMZ (2021) Methylorubrum extorquens DSM 1337.

DSMZ (2021) Paraburkholderia xenovorans DSM 17367.

Farrow, J.M., Wells, G., and Pesci, E.C. (2018) Desiccation tolerance in Acinetobacter baumannii is mediated by the two-component response regulator BfmR, *PLoS One* **13**: e0205638.

Fei, P., Jiang, Y., Feng, J., Forsythe, S.J., Li, R., Zhou, Y., and Man, C. (2017) Antibiotic and Desiccation Resistance of, *Front Microbiol* **8**: 316.

Ferguson, J. (2020) Personal communication on Dec 21, 2020 about growing Pedobacter heparinus DSM 2366. Clipsham, M. (ed).

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data, *Bioinformatics* **28**: 3150-3152.

Garbuz, D. (2017) Regulation of heat shock gene expression in response to stress, *Molecular Biology* **51**: 352-367.

García, A. (2018) Candida sp. As a Potential Reservoir and Transmission Facilitator of Helicobacter Pylori, *Biomedical Journal of Scientific & Technical Research* **4**.

Georlette, D., Damien, B., Blaise, V., Depiereux, E., Uversky, V.N., Gerday, C., and Feller, G. (2003) Structural and functional adaptations to extreme temperatures in psychrophilic, mesophilic, and thermophilic DNA ligases, *J Biol Chem* **278**: 37015-37023.

Ghedira, K., Harigua-Souiai, E., Ben Hamda, C., Fournier, P., Pujic, P., Guesmi, S., et al. (2018) The PEG-responding desiccome of the alder microsymbiont Frankia alni, *Sci Rep* **8**: 759.

Ghosal, D., Omelchenko, M.V., Gaidamakova, E.K., Matrosova, V.Y., Vasilenko, A., Venkateswaran, A., et al. (2005) How radiation kills cells: survival of Deinococcus radiodurans and Shewanella oneidensis under oxidative stress, *FEMS Microbiol Rev* **29**: 361-375.

Golinska, P., Montero-Calasanz, M.D.C., Świecimska, M., Yaramis, A., Igual, J.M., Bull, A.T., and Goodfellow, M. (2020) Modestobacter excelsi sp. nov., a novel actinobacterium isolated from a high altitude Atacama Desert soil, *Syst Appl Microbiol* **43**: 126051.

Gruzdev, N., McClelland, M., Porwollik, S., Ofaim, S., Pinto, R., and Saldinger-Sela, S. (2012) Global transcriptional analysis of dehydrated Salmonella enterica serovar Typhimurium, *Appl Environ Microbiol* **78**: 7866-7875.

Gruzdev, N., Pinto, R., and Sela, S. (2011) Effect of desiccation on tolerance of salmonella enterica to multiple stresses, *Appl Environ Microbiol* **77**: 1667-1673.

Gülez, G., Dechesne, A., Workman, C.T., and Smets, B.F. (2012) Transcriptome dynamics of Pseudomonas putida KT2440 under water stress, *Appl Environ Microbiol* **78**: 676-683.

Hao, Z., Singh, V.P., and Xia, Y. (2018) Seasonal Drought Prediction: Advances, Challenges, and Future Prospects, *Reviews of Geophysics* **56**: 108-141.

Hernández, A., Zamora, J., González, N., Salazar, E., and Sánchez, M.D. (2009) Anhydrobiosis quotient: a novel approach to evaluate stability in desiccated bacterial cells, *J Appl Microbiol* **107**: 436-442.

Hingston, P., Chen, J., Dhillon, B.K., Laing, C., Bertelli, C., Gannon, V., et al. (2017) Genotypes Associated with, Front Microbiol 8: 369.

Hirai, Y. (1991) Survival of bacteria under dry conditions; from a viewpoint of nosocomial infection, *The Journal of hospital infection* **19 3**: 191-200.

Humann, J.L., Ziemkiewicz, H.T., Yurgel, S.N., and Kahn, M.L. (2009) Regulatory and DNA repair genes contribute to the desiccation resistance of Sinorhizobium meliloti Rm1021, Appl Environ Microbiol 75: 446-453.

Jabłońska, J., and Tawfik, D.S. (2019) The number and type of oxygen-utilizing enzymes indicates aerobic vs. anaerobic phenotype, *Free Radic Biol Med* **140**: 84-92.

Janning, B., and in't Veld, P.H. (1994) Susceptibility of bacterial strains to desiccation: a simple method to test their stability in microbiological reference materials, *Analytica Chimica Acta* **286**: 469-476.

Kanehisa, M., and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes, *Nucleic Acids Res* **28**: 27-30.

Kaplan-Levy, R.N., Hadas, O., Summers, M.L., Rucker, J., and Sukenik, A. (2010) Akinetes: Dormant Cells of Cyanobacteria. In: *Dormancy and Resistance in Harsh Environments: Topics in Current Genetics*. Lubzens, E., Cerda, J., and Clark, M. (eds). Berlin, Heidelberg: Springer.

Katoh, H., Asthana, R.K., and Ohmori, M. (2004) Gene expression in the cyanobacterium Anabaena sp. PCC7120 under desiccation, *Microb Ecol* **47**: 164-174.

Khan, A.A., Wang, R.F., Cao, W.W., Franklin, W., and Cerniglia, C.E. (1996) Reclassification of a polycyclic aromatic hydrocarbon-metabolizing bacterium, Beijerinckia sp. strain B1, as Sphingomonas yanoikuyae by fatty acid analysis, protein pattern analysis, DNA-DNA hybridization, and 16S ribosomal DNA sequencing, *Int J Syst Bacteriol* **46**: 466-469.

Kocharunchitt, C., King, T., Gobius, K., Bowman, J.P., and Ross, T. (2012) Integrated transcriptomic and proteomic analysis of the physiological response of Escherichia coli O157:H7 Sakai to steady-state conditions of cold and water activity stress, *Mol Cell Proteomics* **11**: M111.009019.

Kragh, M.L., and Truelstrup Hansen, L. (2019) Initial Transcriptomic Response and Adaption of, *Front Microbiol* **10**: 3132.

Kranner, I., and Birtic, S. (2005) A modulating role for antioxidants in desiccation tolerance, *Integr Comp Biol* **45**: 734-740.

Kriško, A., Smole, Z., Debret, G., Nikolić, N., and Radman, M. (2010) Unstructured hydrophilic sequences in prokaryotic proteomes correlate with dehydration tolerance and host association, *J Mol Biol* **402**: 775-782.

Kuhn, M. (2008) Caret package. Journal of Statistical Software.

Laskowska, E., and Kuczyńska-Wiśnik, D. (2020) New insight into the mechanisms protecting bacteria during desiccation, *Curr Genet* **66**: 313-318.

Le, P.T., Makhalanyane, T.P., Guerrero, L.D., Vikram, S., Van de Peer, Y., and Cowan, D.A. (2016) Comparative Metagenomic Analysis Reveals Mechanisms for Stress Response in Hypoliths from Extreme Hyperarid Deserts, *Genome Biol Evol* **8**: 2737-2747.

LeBlanc, J.C., Gonçalves, E.R., and Mohn, W.W. (2008) Global response to desiccation stress in the soil actinomycete Rhodococcus jostii RHA1, *Appl Environ Microbiol* **74**: 2627-2636.

Lebre, P.H., De Maayer, P., and Cowan, D.A. (2017) Xerotolerant bacteria: surviving through a dry spell, *Nat Rev Microbiol* **15**: 285-296.

Letunic, I., and Bork, P. (2021) Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation, *Nucleic Acids Research* **49**: W293-W296.

Li, H., Bhaskara, A., Megalis, C., and Tortorello, M.L. (2012) Transcriptomic analysis of Salmonella desiccation resistance, *Foodborne Pathog Dis* **9**: 1143-1151.

Ma, Y., Galinski, E.A., Grant, W.D., Oren, A., and Ventosa, A. (2010) Halophiles 2010: life in saline environments, *Appl Environ Microbiol* **76**: 6971-6981.

Macesic, N., Bear Don't Walk, O.J., Pe'er, I., Tatonetti, N.P., Peleg, A.Y., and Uhlemann, A.C. (2020) Predicting Phenotypic Polymyxin Resistance in Klebsiella pneumoniae through Machine Learning Analysis of Genomic Data, *mSystems* **5**.

Madden, T. (2002) The BLAST Sequence Analysis Tool. In: The NCBI Handbook. McEntyre, J., and J, O. (eds). Bethesda (MD): National Center for Biotechnology Information.

Mandal, R.K., and Kwon, Y.M. (2017) Global Screening of, *Front Microbiol* **8**: 1723.

Manzanera, M. (2020) Dealing with water stress and microbial preservation, *Environ Microbiol*.

Martin, M.E., Strachan, R.C., Aranha, H., Evans, S.L., Salin, M.L., Welch, B., et al. (1984) Oxygen toxicity in Streptococcus mutans: manganese, iron, and superoxide dismutase, J Bacteriol 159: 745-749.

Mary, P., Dupuy, N., Dolhem-Biremon, C., Defives, C., and Tailliez, R. (1994) Differences among Rhizobium meliloti and Bradyrhizobium japonicum strains in tolerance to desiccation and storage at different relative humidities, *Soil Biology and Biochemistry* **26**: 1125-1132.

Mauclaire, L., and Egli, M. (2010) Effect of simulated microgravity on growth and production of exopolymeric substances of Micrococcus luteus space and earth isolates, *FEMS Immunol Med Microbiol* **59**: 350-356.

Mitscherlich, E., and Marth, E. (1984) *Microbial survival in the environment: Bacteria and Rickeettsiae important in human and animal health*: Springer-Verlag.

Moradigaravand, D., Palm, M., Farewell, A., Mustonen, V., Warringer, J., and Parts, L. (2018) Prediction of antibiotic resistance in Escherichia coli from large-scale pan-genome data, *PLoS Comput Biol* **14**: e1006258.

National Library of Medicine (US), N.C.f.B.I. (2002) Chapter 18, The Reference Sequence (RefSeq) Project. In: *The NCBI handbook*. Bethesda (MD).

Nishihata, S., Kondo, T., Tanaka, K., Ishikawa, S., Takenaka, S., Kang, C.M., and Yoshida, K.I. (2018) Bradyrhizobium diazoefficiens USDA110 PhaR functions for pleiotropic regulation of cellular processes besides PHB accumulation, *BMC Microbiol* **18**: 156.

Olsson-Francis, K., de la Torre, R., Towner, M.C., and Cockell, C.S. (2009) Survival of akinetes (resting-state cells of cyanobacteria) in low earth orbit and simulated extraterrestrial conditions, *Orig Life Evol Biosph* **39**: 565-579.

Ow, D.S.-W., Nissom, P.M., Philp, R., Oh, S.K.-W., and Yap, M.G.-S. (2006) Global transcriptional analysis of metabolic burden due to plasmid maintenance in Escherichia coli DH5α during batch fermentation, *Enzyme and Microbial Technology* **39**: 391-398.

Paradis, E., and Schliep, K. (2019) ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R, *Bioinformatics* **35**: 526-528.

Potts, M. (1994) Desiccation tolerance of prokaryotes, *Microbiol Rev* **58**: 755-805.

Potts, M., Slaughter, S.M., Hunneke, F.U., Garst, J.F., and Helm, R.F. (2005) Desiccation tolerance of prokaryotes: application of principles to human cells, *Integr Comp Biol* **45**: 800-809.

Prasad, J., McJarrow, P., and Gopal, P. (2003) Heat and osmotic stress responses of probiotic Lactobacillus rhamnosus HN001 (DR20) in relation to viability after drying, *Appl Environ Microbiol* **69**: 917-925.

Prasad, R., and Aranda, E. (2019) *Approaches in Bioremediation: The New Era of Environmental Microbiology and Nanobiotechnology*: Springer Nature.

Reichenbach, H., and Dworkin, M. (1992) The Myxobacteria. In: *The Prokaryotes*. Balows, A., Truper, H.G., Dworkin, M., Harder, W., and Schleifer, K.H. (eds). New York, NY: Springer.

River, C. (2011) Technical Sheet: Helicobacter species. International, C.R.L. (ed).

Robinson, S.L., Smith, M.D., Richman, J.E., Aukema, K.G., and Wackett, L.P. (2020) Machine learning-based prediction of activity and substrate specificity for OleA enzymes in the thiolase superfamily, *Synthetic Biology* **5**.

Rodriguez-Salazar, J., Moreno, S., and Espín, G. (2017) LEA proteins are involved in cyst desiccation resistance and other abiotic stresses in Azotobacter vinelandii, *Cell Stress Chaperones* **22**: 397-408.

Rogozin, I.B., Makarova, K.S., Natale, D.A., Spiridonov, A.N., Tatusov, R.L., Wolf, Y.I., et al. (2002) Congruent evolution of different classes of non-coding DNA in prokaryotic genomes, *Nucleic Acids Res* **30**: 4264-4271.

Romanovskaya, V., Rokitko, P., Mikheev, A., Gushcha, N., Malashenko, Y., and Chernaya, N. (2002) The Effect of γ-Radiation and Desiccation on the Viability of the Soil Bacteria Isolated from the Alienated Zone around the Chernobyl Nuclear Power Plant, *Microbiology* **71**: 608-613.

Ryabova, A., Kozlova, O., Kadirov, A., Ananeva, A., Gusev, O., and Shagimardanova, E. (2020) DetR DB: A Database of Ionizing Radiation Resistance Determinants, *Genes (Basel)* **11**.

Sadasivan, L., and Neyra, C.A. (1985) Flocculation in Azospirillum brasilense and Azospirillum lipoferum: exopolysaccharides and cyst formation, *J Bacteriol* **163**: 716-723.

SantaCruz-Calvo, L., González-López, J., and Manzanera, M. (2013) Arthrobacter siccitolerans sp. nov., a highly desiccation-tolerant, xeroprotectant-producing strain isolated from dry soil, *Int J Syst Evol Microbiol* **63**: 4174-4180.

Shams, A.M., Rose, L.J., Hodges, L., and Arduino, M.J. (2007) Survival of Burkholderia pseudomallei on Environmental Surfaces, *Appl Environ Microbiol* **73**: 8001-8004.

Shang, J.L., Zhang, Z.C., Yin, X.Y., Chen, M., Hao, F.H., Wang, K., et al. (2018) UV-B induced biosynthesis of a novel sunscreen compound in solar radiation and desiccation tolerant cyanobacteria, *Environ Microbiol* **20**: 200-213.

Sharma, G., Narwani, T., and Subramanian, S. (2016) Complete Genome Sequence and Comparative Genomics of a Novel Myxobacterium Myxococcus hansupus, *PLoS One* **11**: e0148593.

Shirkey, B., Kovarcik, D.P., Wright, D.J., Wilmoth, G., Prickett, T.F., Helm, R.F., et al. (2000) Active Fe-containing superoxide dismutase and abundant sodF mRNA in Nostoc commune (Cyanobacteria) after years of desiccation, *J Bacteriol* **182**: 189-197.

Shukla, M., Chaturvedi, R., Tamhane, D., Vyas, P., Archana, G., Apte, S., et al. (2007) Multiple-stress tolerance of ionizing radiation-resistant bacterial isolates obtained from various habitats: correlation between stresses, *Curr Microbiol* **54**: 142-148.

Singh, J., Kumar, D., Ramakrishnan, N., Singhal, V., Jervis, J., Garst, J.F., et al. (2005) Transcriptional response of Saccharomyces cerevisiae to desiccation and rehydration, *Appl Environ Microbiol* **71**: 8752-8763.

Smith, S., Meade, J., Gibbons, J., McGill, K., Bolton, D., and Whyte, P. (2016) The impact of environmental conditions on Campylobacter jejuni survival in broiler faeces and litter, *Infect Ecol Epidemiol* **6**: 31685.

Soparkar, M.B. (1917) The Vitality of the Tubercle Bacillus outside the Body., *Indian Journal of Medical Research* **4**: 627-650.

Srikumar, S., Cao, Y., Yan, Q., Van Hoorde, K., Nguyen, S., Cooney, S., et al. (2019) RNA Sequencing-Based Transcriptional Overview of Xerotolerance in Cronobacter sakazakii SP291, *Appl Environ Microbiol* **85**.

Su, M., Satola, S.W., and Read, T.D. (2019) Genome-Based Prediction of Bacterial Antibiotic Resistance, *J Clin Microbiol* **57**.

Tassoulas, L.J., Robinson, A., Martinez-Vaz, B., Aukema, K.G., and Wackett, L.P. (2021) Filling in the Gaps in Metformin Biodegradation: a New Enzyme and a Metabolic Pathway for Guanylurea, *Appl Environ Microbiol* **87**.

Ujaoney, A.K., Padwal, M.K., and Basu, B. (2017) Proteome dynamics during post-desiccation recovery reveal convergence of desiccation and gamma

radiation stress response pathways in Deinococcus radiodurans, *Biochim Biophys Acta Proteins Proteom* **1865**: 1215-1226.

Ulrich, N., Nagler, K., Laue, M., Cockell, C.S., Setlow, P., and Moeller, R. (2018) Experimental studies addressing the longevity of Bacillus subtilis spores - The first data from a 500-year experiment, *PLoS One* **13**: e0208425.

Vriezen, J.A., de Bruijn, F.J., and Nüsslein, K. (2006) Desiccation responses and survival of Sinorhizobium meliloti USDA 1021 in relation to growth phase, temperature, chloride and sulfate availability, *Lett Appl Microbiol* **42**: 172-178.

Vriezen, J.A., de Bruijn, F.J., and Nüsslein, K.R. (2012) Desiccation induces viable but Non-Culturable cells in Sinorhizobium meliloti 1021, *AMB Express* **2**: 6.

Vázquez-Boland, J.A., and Meijer, W.G. (2019) The pathogenic actinobacterium Rhodococcus equi: what's in a name?, *Mol Microbiol* **112**: 1-15.

Wilkinson, T.R. (1966) Survival of bacteria on metal surfaces, *Appl Microbiol* **14**: 303-307.

Wright, M., and Ziegler, A. (2017) A Fast Implementation of Random Forests for High Dimensional Data in C++ and R, Journal of Statistical Software 77: 1-17.

Yadav, A.N., Kour, D., Kaur, T., Devi, R., Guleria, G., Rana, K.L., et al. (2020) Chapter 18 - Microbial biotechnology for sustainable biomedicine systems: Current research and future challenges. In: *New and Future Developments in Microbial Biotechnology and Bioengineering*. Rastegari, A.A., Yadav, A.N., and Yadav, N. (eds): Elsevier. 281-292.

Zeidler, S., and Müller, V. (2019) The role of compatible solutes in desiccation resistance of Acinetobacter baumannii, *Microbiologyopen* **8**: e00740.

Zhang, X., Al-Dossary, A., Hussain, M., Setlow, P., and Li, J. (2020) Applications of Bacillus subtilis Spores in Biotechnology and Advanced Materials, *Appl Environ Microbiol* **86**.

Appendix A: Supplemental Materials

Table A1: Desiccation tolerant bacteria in the model.

Species	RefSeq Accession number	Assigned desiccation status	Reference
Nostoc commune HK-02 NIES-4070	GCA_003113895.1	Tolerant	Shirkey et al., 2000
Deinococcus radiodurans R1 dM1	GCA_008329785.1	Tolerant	Ujaoney et al., 2017
Bacillus subtilis NCIB 3610	GCA_006088795.1	Tolerant	Ulrich et al., 2018
Listeria monocytogenes 08-5578	GCA_000093125.2	Tolerant	Kragh et al., 2019
Salmonella Typhimurium 14028s	GCA_006088735.1	Tolerant	Chen et al., 2018
Salmonella Typhimurium 14028s	GCA_003864015.1	Tolerant	Chen et al., 2018
Cronobacter sakazakii SP291	GCA_000339015.1	Tolerant	Srikumar et al., 2019
Sinorhizobium meliloti USDA1021	GCA_002197445.1	Tolerant	Vriezen et al., 2012
Anabaena cylindrica PCC 7122	GCA_000317695.1	Tolerant	Olsson-Francis et al., 2009
Nostoc flagelliforme CCNUN1	GCA_002813575.1	Tolerant	Shang et al., 2018
Chroococcidiopsis thermalis PCC 7203	GCA_000317125.1	Tolerant	Potts, 1994
Gloeobacter violaceus PCC 7421	GCA_000011385.1	Tolerant	Potts et al., 2005
Staphylococcus aureus NCTC13811	GCA_900637155.1	Tolerant	Chaibenjawong et al., 2011
Enterococcus faecium strain 4928STDY7387800	GCA_902166835.1	Tolerant	Janning et al., 1994
Salmonella enterica subsp. enterica serovar Panama str. ATCC 7378	GCA_000486765.2	Tolerant	Janning et al., 1994
Methylorubrum extorquens strain TK 0001	GCA_900234795.1	Tolerant	Romanovskaya et al., 2002
Methylobacterium mesophilicum SR1.6/6	GCA_000364445.2	Tolerant	Romanovskaya et al., 2002
Nocardia asteroides strain NCTC11293	GCA_900637185.1	Tolerant	Mitscherlich et al., 1984
Cronobacter sakazakii ATCC 29544	GCA_000982825.1	Tolerant	Fei et al., 2017
Tsukamurella paurometabola DSM 20162	GCA_000092225.1	Tolerant	Hernandez et al., 2009
Modestobacter marinus strain BC501	GCA_000306785.1	Tolerant	Golinska et al., 2020
Myxococcus stipitatus DSM 14675	GCA_000331735.1	Tolerant	Sharma et al., 016
Rhodococcus hoagii 103S	GCA_000196695.1	Tolerant	Vazquez-Boland et al., 2019
Clostridium tetani E88	GCA_000007625.1	Tolerant	Zeidler et al., 2019
Streptomyces coelicolor A3(2) strain CFB_NBC_0001	GCA_008931305.1	Tolerant	Chater, 2001
Bradyrhizobium japonicum USDA 6	GCA_000284375.1	Tolerant	Antheunisse, et al., 1981
Azotobacter vinelandii DJ	GCA_000021045.1	Tolerant	Antheunisse, et al., 1981
Azospirillum brasilense strain Sp 7	GCA_001315015.1	Tolerant	Sadasivan et al., 1985
Rhizobium leguminosarum strain Vaf-108	GCA_001890425.1	Tolerant	Antheunisse, et al., 1981
Mycobacterium Leprae strain MRHRU-235-G	GCA_003253775.1	Tolerant	Desikan et al., 1995
Mycobacterium tuberculosis strain DKC2	GCA_900520315.1	Tolerant	Soparkar, 1917
Arthrobacter crystallopoietes strain DSM 20117	GCA_002849715.1	Tolerant	Boylen, 1973
Lactobacillus rhamnosus GG ATCC 53103	GCA_000011045.1	Tolerant	Prasad et al., 2003
Lactobacillus rhamnosus GG	GCA_003353455.1	Tolerant	Prasad et al., 2003
Corynebacterium diphtheriae strain NCTC11397	GCA_001457455.1	Tolerant	Mitscherlich, et al., 1984

Table A2: Desiccation sensitive bacteria in the model.

Species	RefSeq Accession number	Assigned desiccation status	Reference
E. coli MG1655	GCA_000005845.2	Sensitive	Chen, et al., 2018
E. coli DH10B	GCA_000019425.1	Sensitive	Potts et al., 2005
Treponema pallidum subsp. Pallidum (strain X-4)	GCA_005885795.1	Sensitive	Potts, 1994
Vibrio cholerae strain NCTC 30	GCA_900538065.1	Sensitive	Potts, 1994
Leptospira interrogans serovar Icterohaemorrhagiae	GCA_001683775.2	Sensitive	Potts, 1994
Aeromonas hydrophila strain NEB724	GCA_012273595.1	Sensitive	Janning, et al., 1994
Pseudomonas aeruginosa strain C7-25	GCA_902703215.1	Sensitive	Janning, et al., 1994
Neisseria meningitidis strain NCTC10026	GCA_900638605.1	Sensitive	Potts, 1994
Klebsiella pneumoniae Kpn2166	GCA_902723705.1	Sensitive	Potts, 1994
Pasteurella multocida strain NCTC10323	GCA_900638665.1	Sensitive	Mitscherlich, et al., 1984
Neisseria gonorrhoeae strain NCTC13484	GCA_900637245.1	Sensitive	Mitscherlich, et al., 1984
Ehrlichia ruminantium strain Kumm2	GCA_009728855.1	Sensitive	Mitscherlich, et al., 1984
Streptococcus salivarius strain NCTC8618	GCA_900636435.1	Sensitive	Mitscherlich, et al., 1984
Moraxella bovis strain Epp63	GCA_003287015.1	Sensitive	Mitscherlich, et al., 1984
Mycoplasma bovis JF4278	GCA_900088685.1	Sensitive	Mitscherlich, et al., 1984
Serratia marcescens strain 4928STDY7387938	GCA_902166755.1	Sensitive	Mitscherlich, et al., 1984
Yersinia pestis strain SCPM-O-B-5942 (I-2638)	GCA_009363195.1	Sensitive	Mitscherlich, et al., 1984
Campylobacter jejuni subsp. Jejuni strain NCTC10983	GCA_900638365.1	Sensitive	Smith et al., 2016
Acinetobacter baumannii strain ATCC 17978	GCA_011067065.1	Sensitive	Farrow et al., 2018
Acinetobacter baumannii strain ATCC 19606	GCA_009759685.1	Sensitive	Farrow et al., 2018
Helicobacter bilis strain AAQJH	GCA_001999985.1	Sensitive	Charles River, 2011
Helicobacter pylori ATCC 4350	GCA_900478295.1	Sensitive	Gracia, 2018
Enterobacter cloacae subsp. cloacae ATCC 13047	GCA_000025565.1	Sensitive	Janning, et al., 1994
Pseudomonas putida strain KT2440	GCA_900167985.1	Sensitive	SantaCruz-Calvo et al., 2013
Shewanella oneidensis MR-1	GCA_000146165.2	Sensitive	Daly et al., 2004
Bordetella pertussis 18323	GCA_000306945.1	Sensitive	Mitscherlich, et al., 1984
Streptococcus pneumoniae strain NCTC7465	GCA_001457635.1	Sensitive	Mitscherlich, et al., 1984
Haemophilus influenzae strain NCTC8143	GCA_001457655.1	Sensitive	Mitscherlich, et al., 1984
Francisella tularensis subsp. novicida U112	GCA_000833375.1	Sensitive	Wilkinson, 1966
Burkholderia pseudomallei MSHR2543	GCA_000959225.1	Sensitive	Shams et al., 2007
Rickettsia rickettsii str. Iowa	GCA_000017445.3	Sensitive	Mitscherlich, et al., 1984
Brucella abortus 2308	GCA_000054005.1	Sensitive	Mitscherlich, et al., 1984
Legionella pneumophila subsp. Pascullei strain NCTC12273	GCA_900637585.1	Sensitive	Mitscherlich, et al., 1984

Figure A1: Classification accuracies of 100 random 75% train- 25% test splits. (A) Classification accuracies of the full gene counts model, (B) Classification accuracies of the top 30 most important genes from the gene counts model, (C) classification accuracies of the full binary model, and (D) classification accuracies of the top 30 most important genes from the binary model.

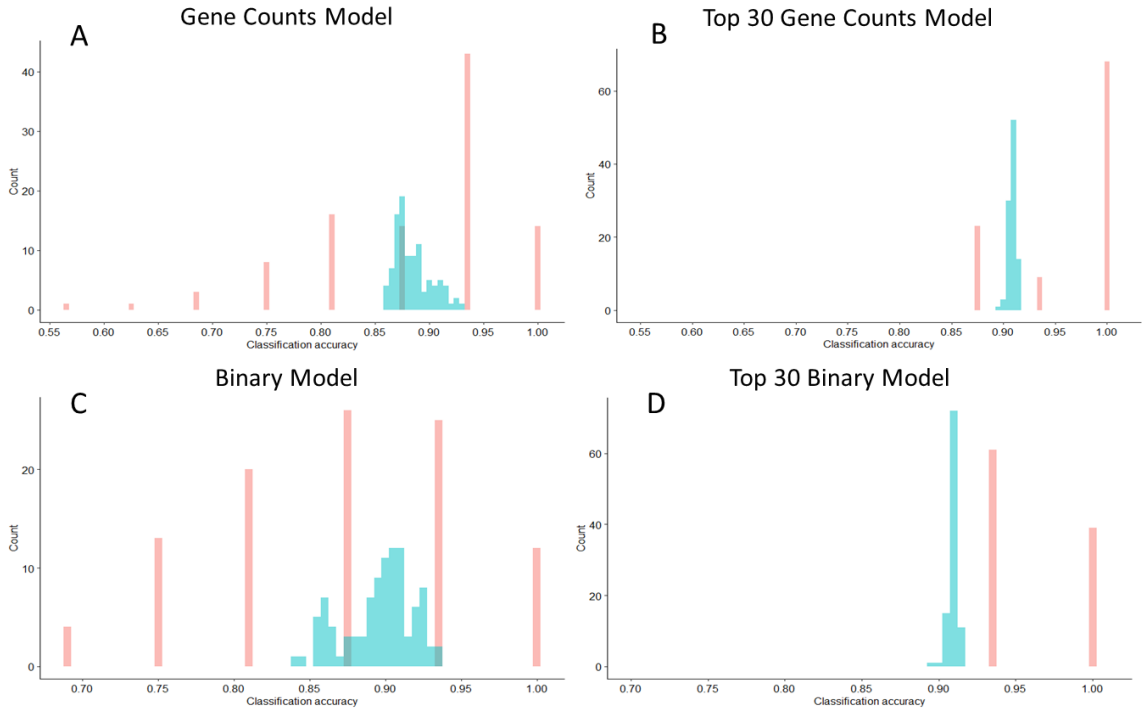

Table A3: Example variable importance plot showing the top 30 most important features in the gene counts model given in the order of importance. Bolded genes are shared with the top 30 binary genes.

	Genes	Protein Name/description	Protein Function	Organism(s) or origin	Tolerant average	Sensitive Average	tol:sen ratio
1	Q8ZRS8	Aconitate hydratase B acnB	RNA binding	Salmonella typhimurium (strain LT2 / SGSC1412 / ATCC 700720) & Cronobacter sakazakii SP291	0.31	0.64	0.49
2	**AGE84778.1**	30S ribosomal protein S6	Ribosomal protein	Cronobacter sakazakii SP291	0.17	0.67	0.26
3	AGE87481.1	tRNA (Thr-GGU) A37 N-methylase	Methylation	Cronobacter sakazakii SP291	0.17	0.55	0.31
4	**Q7CQ12**	Iron-binding protein IscA iscA	Iron binding	Salmonella typhimurium (strain LT2 / SGSC1412 / ATCC 700720) & Cronobacter sakazakii SP291	0.34	0.82	0.42
5	**ABG93071.1**	heat-inducible transcription repressor HrcA	Stress response	Rhodococcus jostii RHA1	0.80	0.12	6.60
6	**AGE87619.1**	diadenosine tetraphosphatase	Hydrolase	Cronobacter sakazakii SP291	0.17	0.67	0.26
7	Q8ZNG4	Anaerobic glycerol-3-phosphate dehydrogenase subunit B glpB	Glycerol catabolic process	Salmonella typhimurium (strain LT2 / SGSC1412 / ATCC 700720)	0.09	0.30	0.28
8	**P0DJM4**	Heat-inducible transcription repressor HrcA	Repressor	Salmonella typhimurium (strain LT2 / SGSC1412 / ATCC 700720)	0.80	0.21	3.77
9	**ABG97784.1**	short chain dehydrogenase	Oxidoreductase	Rhodococcus jostii RHA1	2.89	0.76	3.81
10	ABG92748.1	short chain dehydrogenase	Oxidoreductase	Rhodococcus jostii RHA1	5.89	3.61	1.63
11	CAC99332.1	alpha,alpha-phosphotrehalase	Trehalose catabolic process	Listeria monocytogenes EGD-e	2.23	0.64	3.50
12	WP_011607074.1	gamma-aminobutyraldehyde dehydrogenase	Oxidoreductase	Frankia alni	7.03	5.45	1.29
13	AGE85424.1	aldo-keto reductase	Ascorbate biosynthesis	Cronobacter sakazakii SP291	3.91	1.73	2.27
14	ABG92284.1	60 kDa chaperonin GroEL	Protein refolding	Rhodococcus jostii RHA1	1.83	1.00	1.83
15	**AGE84780.1**	30S ribosomal protein S18	Ribosomal protein	Cronobacter sakazakii SP291	0.17	0.67	0.26
16	CAC99944.1	Kinase/pyrophosphorylase	Kinase	Listeria monocytogenes EGD-e	0.40	0.09	4.40
17	AGE86015.1	tRNA threonylcarbamoyl adenosine modification protein	tRNA processing	Cronobacter sakazakii SP291	0.17	0.52	0.33
18	AGE87125.1	D-alanyl-D-alanine carboxypeptidase fraction A	Cell wall biogenesis	Cronobacter sakazakii SP291	0.89	1.24	0.71
19	AGE87878.1	ribosome maturation protein RimP	Ribosome biogenesis	Cronobacter sakazakii SP291	0.26	0.64	0.40
20	ABG95190.1	FAD-binding oxidoreductase	Oxidoreductase	Rhodococcus jostii RHA1	2.20	1.18	1.86
21	WP_000824854.1	ABC transporter permease	Transmembrane transport	Salmonella typhimurium (strain LT2 / SGSC1412 / ATCC 700720)	1.26	0.45	2.77
22	AGE85681.1	protease	Protease	Cronobacter sakazakii SP291	0.74	1.12	0.66
23	**AGE88559.1**	ribosome-associated heat shock protein Hsp15	Stress response	Cronobacter sakazakii SP291	0.17	0.58	0.30
24	AGE85947.1	Holliday junction resolvase	DNA repair	Cronobacter sakazakii SP291	0.37	0.79	0.47
25	ABG96218.1	possible dithiol-disulfide isomerase	Polyketide synthesis	Rhodococcus jostii RHA1	0.40	0.12	3.30
26	**AGE85293.1**	phosphoribosylformylglycinamidine synthase	Purine biosynthesis	Cronobacter sakazakii SP291	0.26	0.70	0.37
27	P06173	DNA-directed RNA polymerase subunit beta rpoB	Transcription	Salmonella typhimurium (strain LT2 / SGSC1412 / ATCC 700720) & Cronobacter sakazakii SP291	2.14	1.18	1.81
28	**P66451**	30S ribosomal protein S17 rpsQ	Ribosomal protein	Salmonella typhimurium (strain LT2 / SGSC1412 / ATCC 700720)	0.20	0.61	0.33
29	ABG99104.1	biphenyl 2,3-dioxygenase, reductase	aromatic hydrocarbon catabolism	Rhodococcus jostii RHA1	1.31	0.52	2.55
30	AGE87285.1	ammonium transporter	Ammonia transport	Cronobacter sakazakii SP291	1.26	0.82	1.54

Table A4: Example variable importance plot showing the top 30 most important features in the binary model given in the order of importance. Bolded genes are shared with the top 30 gene counts genes.

	Gene	Protein Name/Description	Protein Function	organism(s) of origin	Tolerant average	Sensitive Average	Ratio Tol:sen
1	**ABG93071.1**	heat-inducible transcription repressor HrcA	Stress response	Rhodococcus jostii RHA1	0.80	0.12	6.60
2	CAC99454.1	lmo1376 (6-phosphogluconate dehydrogenase, decarboxylating)	Gluconate utilization	Listeria monocytogenes EGD-e	0.97	0.52	1.89
3	AGE84907.1	hypothetical protein CSSP291_01545	Carotenoid biosynthesis	Cronobacter sakazakii SP291	0.57	0.00	20.00
4	Q9AGE7	10 kDa chaperonin	Chaperone	Listeria monocytogenes 08-5578	0.74	0.18	4.09
5	**AGE87619.1**	symmetrical bis(5'-nucleosyl)-tetraphosphatase	hydrolase	Cronobacter sakazakii SP291	0.17	0.67	0.26
6	ABG98962.1	FeS assembly ATPase	protein turnover	Rhodococcus jostii RHA1	0.97	0.42	2.29
7	CAC98691.1	lmo0613	oxidoreductase	Listeria monocytogenes EGD-e	0.97	0.58	1.69
8	**AGE84780.1**	30S ribosomal protein S18	Ribosomal Protein	Cronobacter sakazakii SP291	0.17	0.67	0.26
9	**P0DJM4**	Heat-inducible transcription repressor HrcA	Stress response	Listeria monocytogenes 08-5578	0.80	0.21	3.77
10	CAD00344.1	lmo2266	secondary metabolism	Listeria monocytogenes EGD-e	0.91	0.39	2.32
11	AGE87210.1	UDP-2,3-diacylglucosamine hydrolase	lipid biosynthesis/metabolism	Cronobacter sakazakii SP291	0.17	0.67	0.26
12	**AGE84778.1**	30S ribosomal protein S6	Ribosomal Protein	Cronobacter sakazakii SP291	0.17	0.67	0.26
13	AGE88280.1	F0F1 ATP synthase subunit B	ATP synthesis	Cronobacter sakazakii SP291	0.17	0.67	0.26
14	**ABG97784.1**	short chain dehydrogenase	oxidoreductase	Rhodococcus jostii RHA1	0.80	0.27	2.93
15	AGE87926.1	stringent starvation protein A	glutathion metabolic process	Cronobacter sakazakii SP291	0.17	0.64	0.27
16	ABG98960.1	FeS assembly protein	iron-sulfur cluster assembly	Rhodococcus jostii RHA1	0.91	0.42	2.16
17	AGE87876.1	translation initiation factor IF-2	protein biosynthesis	Cronobacter sakazakii SP291	0.23	0.67	0.34
18	P0A297	50S ribosomal protein L10 rplJ	Ribosomal Protein	Salmonella typhimurium (strain LT2 / SGSC1412 / ATCC 700720) & Cronobacter sakazakii SP291	0.23	0.67	0.34
19	**AGE85293.1**	phosphoribosylformylglycinamidine synthase	purine biosynthesis	Cronobacter sakazakii SP291	0.26	0.70	0.37
20	CAD00083.1	lmo2005	D-threo-aldose 1-dehydrogenase activity	Listeria monocytogenes EGD-e	0.97	0.61	1.60
21	CAD00201.1	lmo2123	maltose transport	Listeria monocytogenes EGD-e	0.86	0.45	1.89
22	AGE85244.1	16S rRNA-processing protein RimM	ribosome biogenesis	Cronobacter sakazakii SP291	0.17	0.58	0.30
23	CAC98837.1	lmo0759	vicinal oxygen chelate (VOC) family	Listeria monocytogenes EGD-e	0.34	0.00	20.00
24	ABG93921.1	probable glycosyltransferase	Glycosyltransferase	Rhodococcus jostii RHA1	0.40	0.03	13.20
25	WP_041938850.1	50S ribosomal protein L18	Ribosomal Protein	Frankia alni	0.80	0.21	3.77
26	**P66451**	30S ribosomal protein S17 rpsQ	Ribosomal Protein	Salmonella typhimurium (strain LT2 / SGSC1412 / ATCC 700720)	0.20	0.61	0.33
27	**Q7CQ12**	Iron-binding protein IscA	iron-sulfur cluster assembly	Salmonella typhimurium (strain LT2 / SGSC1412 / ATCC 700720) & Cronobacter sakazakii SP291	0.20	0.64	0.31
28	**AGE88559.1**	ribosome-associated heat shock protein Hsp15	Stress response	Cronobacter sakazakii SP291	0.17	0.58	0.30
29	AGE88279.1	F0F1 ATP synthase subunit delta	ATP synthesis	Cronobacter sakazakii SP291	0.17	0.55	0.31
30	ABG98974.1	probable transcriptional regulator, MoxR family protein	Transcription regulator	Rhodococcus jostii RHA1	0.77	0.33	2.31

Figure A2: Phylogeny tree of all the bacterial species in the model and the experimental assay. The species that are only present in the model are in dark red text, the species in both the model and the experiment are in black text, and the species only in the experiment are in blue text. The tree scale showing genetic distance is shown in the upper left-hand corner.

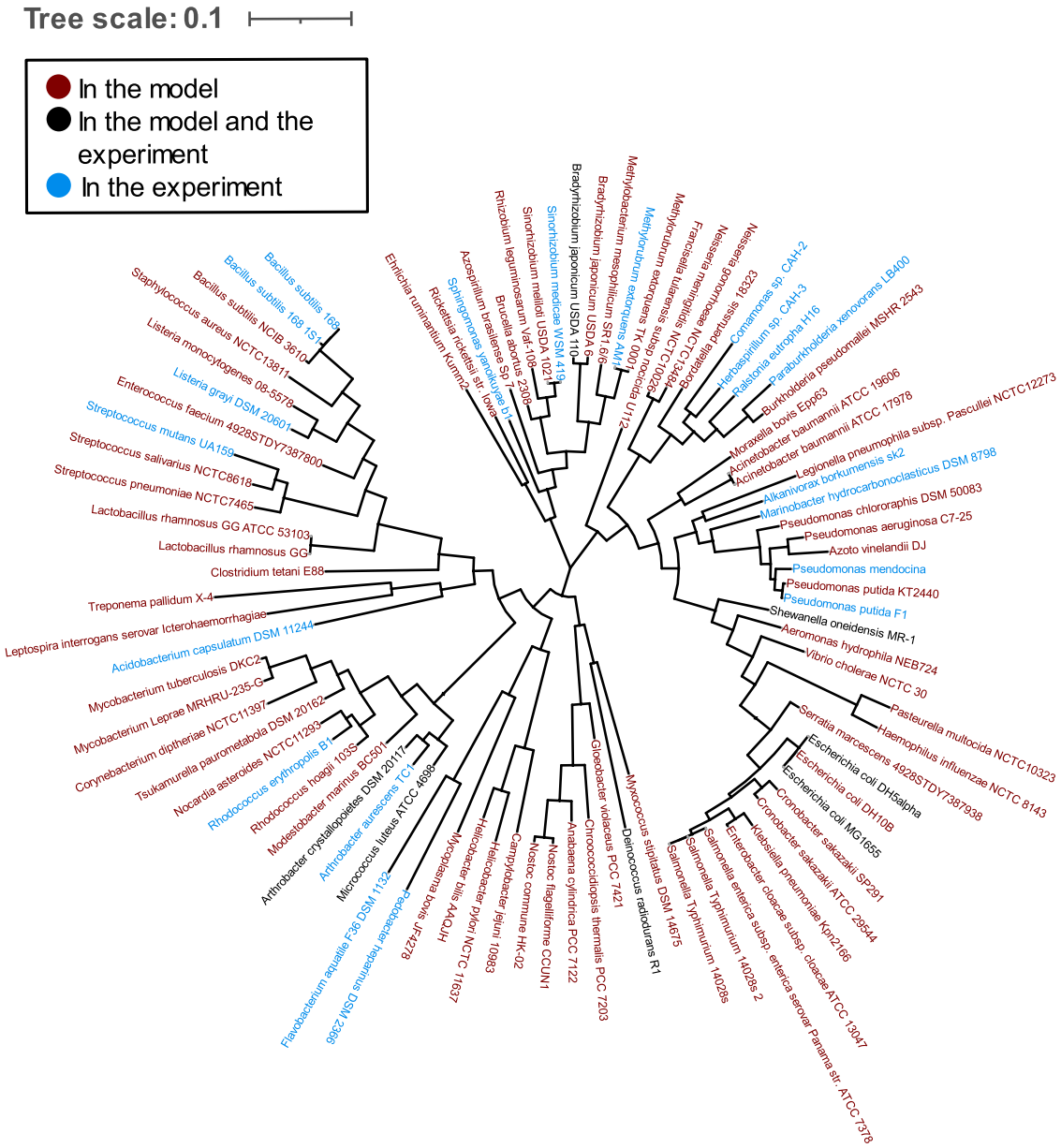

Table A5: The predictions from all the models for the bacteria used in the experimental model validation, including the full gene counts model, the top 30 gene counts model, the full binary model, and the top 30 binary model. Prediction scores >0.5 indicate the species is predicted to be desiccation tolerant (green), scores <0.5 are predicted to be desiccation sensitive. Scores closer to 0 or 1 indicate the model has a higher confidence in the prediction. Scores closer to 0.5 indicate the model is less confident in the prediction. Most of the predictions remained similar for the bacteria across the models. Only two bacteria, the *Pedobacter heparinus* DSM 2366 and *Paraburkholderia xenovorans* LB400 switched from being predicted desiccation sensitive to desiccation tolerant.

Bacterial Strains	Full gene counts	Top 30 gene counts	Full Binary	Top 30 Binary
Arthrobacter crystallopoietes DSM 20117	0.95	1.00	0.96	0.98
Rhodococcus B1	0.95	0.99	0.96	0.99
Bacillus subtilis 168 WT vegetative cells	0.94	0.98	0.95	0.93
Bacillus subtilis 168 1S1 Non-spore mutant	0.94	0.98	0.95	0.93
Arthrobacter aurescens TC1	0.90	1.00	0.91	0.98
Bradyrhizobium diazoefficiens USDA 110	0.89	0.97	0.88	1.00
Listeria grayi DSM 20601	0.89	0.99	0.94	0.97
WSM 419 *Sinorhizobium medicae*	0.87	0.95	0.89	0.85
Deinococcus radiodurans R1	0.85	0.86	0.87	0.82
Methylobacterium extorquens AM1	0.85	0.94	0.86	0.99
Micrococcus luteus ATCC 4698	0.71	0.82	0.81	0.80
Sphingomonas yanoikuae b1	0.58	0.67	0.63	0.58
Acidobacterium capsulatum DSM 11244	0.54	0.54	0.58	0.55
Pedobacter heparinus DSM 2366	0.48	0.75	0.57	0.76
Paraburkholderia xenovorans LB400	0.47	0.40	0.38	0.51
Ralstonia eutropha H16	0.46	0.38	0.35	0.26
Herbaspirillum sp. CAH-3	0.40	0.23	0.35	0.30
Streptococcus mutans UA159	0.36	0.21	0.36	0.06
Flavobacterium aquatile F36	0.35	0.18	0.39	0.25
Pseudomonas mendocina	0.34	0.23	0.28	0.48
Comamonas sp. CAH-2	0.31	0.27	0.26	0.27
Marinobacter hydrocarbonoclasticus DSM 8798	0.30	0.08	0.17	0.13
Pseudomonas chlororaphis DSM 50083	0.27	0.09	0.19	0.22
Pseudomonas putida F1	0.20	0.16	0.14	0.31
Alkanivorax borkumensis sk2	0.20	0.06	0.20	0.15
Escherichia coli DH5alpha	0.14	0.11	0.18	0.19
Escherichia coli MG1655	0.09	0.11	0.16	0.19
Shewanella oneidensis MR-1	0.07	0.01	0.06	0.02

Table A6: The strains used in the experimental validation, the source, Refseq accession number of the genome, the media for cultivation, and the temperature at which they were grown. The broth was supplemented with 1.5% (w/v) of agar for solid media. *Streptococcus mutans* UA159 was grown anaerobically.

Bacterial Strains	Strain or Source	Refseq Accession Number	Media	Reference	Temperature (°C)
Rhodococcus B1	Laboratory Isolate	Unpublished genome	BD Difco™ LB Broth	Laboratory Isolate	28
Pseudomonas mendocina	Laboratory Isolate	Unpublished genome	BD Difco™ LB Broth	Tassoulas et al., 2021	28
Comamonas sp. CAH-2	Laboratory Isolate	GCA_009668305.1	BD Difco™ LB Broth	Laboratory Isolate	28
Escherichia coli DH5alpha	DSM 6897	GCA_002848225.1	BD Difco™ LB Broth	Ow et al., 2006	28
Escherichia coli MG1655	DSM 18039	GCA_000005845.2	BD Difco™ LB Broth	Baev et al., 2006	28
Arthrobacter crystallopoietes	*ATCC 15481*	GCA_002849715.1	Oxoid Nutrient Broth	ATCC 15481, 2021	28
Arthrobacter aurescens TC1	Gift from Lab of Dr. Sadowsky at UMN	GCA_000014925.1	Oxoid Nutrient Broth	ATCC 13344, 2021	28
Ralstonia eutropha H16	DSM 428	GCA_004798725.1	Oxoid Nutrient Broth	ATCC 17699, 2021	28
Herbaspirillum sp. CAH-3	Laboratory Isolate	GCA_009668275.1	Oxoid Nutrient Broth	Laboratory Isolate	28
Pseudomonas chlororaphis	DSM 55083	GCA_003945765.1	Oxoid Nutrient Broth	ATCC 9446, 2021	28
Pseudomonas putida F1	DSM 6899	GCA_000016865.1	Oxoid Nutrient Broth	ATCC 700007, 2021	28
Sinorhizobium medicae WSM 419	Gift from Lab of Dr. Sadowsky at UMN	GCA_000017145.1	TY Medium (DSMZ Medium 1143)	Bailly et al., 2011	28
Bradyrhizobium USDA 110	NRRL B-4361	GCA_000011365.1	TY Medium (DSMZ Medium 1143)	Nishihata et al., 2018	28
Flavobacterium aquatile F36	DSM 1132	GCA_002217235.1	R2A Medium (DSMZ Medium 830)	DSMZ 1132, 2021	28
Bacillus subtilis 168 WT vegetative cells	DSM 23778	GCA_013009385.1	BD Difco™ Tryptic Soy Broth	DSMZ 23778, 2021	28
Bacillus subtilis 168 1S1	BGSC 1S1		BD Difco™ Tryptic Soy Broth	DSMZ 23778, 2021	28
Deinococcus radiodurans R1	DSM 20539	GCA_008329785.1	BD Difco™ Tryptic Soy Broth	Bacdive 3852, 2021	28
Micrococcus luteus	ATCC 4698	GCA_900475555.1	BD Difco™ Tryptic Soy Broth	ATCC 4698, 2021	28
Sphingomonas yanoikuae b1	DSM 6900	GCA_000731935.1	BD Difco™ Tryptic Soy Broth	Kahn et al., 1996	28
Pedobacter heparinus DSM 2366	Gift from Lab of Dr. Sadowsky at UMN	GCA_000023825.1	BD Difco™ Tryptic Soy Broth	Ferguson, 2020	28
Paraburkholderia xenovorans LB400	DSM 17367	GCA_000756045.1	BD Difco™ Tryptic Soy Broth	DSMZ 17367, 2021	28
Shewanella oneidensis MR-1	ATCC 700550	GCA_000146165.2	BD Difco™ Tryptic Soy Broth	ATCC 700550, 2021	28
Acidobacterium capsulatum	DSM 11244	GCA_000022565.1	Acidophilum Medium (DSMZ Medium 269), with strain-specific modifications (pH 3-6)	DSMZ 11244, 2021	28
Marinobacter hydrocarbonoclasticus	DSM 8798	GCA_000284615.1	Zobell Marine Broth 2216	ATCC 49840, 2021	28
Alkanivorax borkumensis sk2	DSM 11573	GCA_000009365.1	Zobell marine broth 2216 with 10g/L Na pyruvate	DSMZ 11573, 2021	28
Listeria grayi	DSMZ 20601	GCA_000148995.1	BD Difco™ Brain Heart Infusion Broth	DSMZ 20601, 2021	37
Streptococcus mutans UA159	ATCC 700610	GCA_000007465.2	BD Difco™ Brain Heart Infusion Broth	Ahn et al., 2005	37
Methylobacterium extorquens AM1	DSM 1338	GCA_000021845.1	Oxoid nutrient broth with 1% (v/v) methanol	DSMZ 1338, 2021	28

Figure A3: Normalized viability at 13 weeks. Viability was normalized by dividing the 13 week Log (CFU/mL) by the initial Log (CFU/mL).

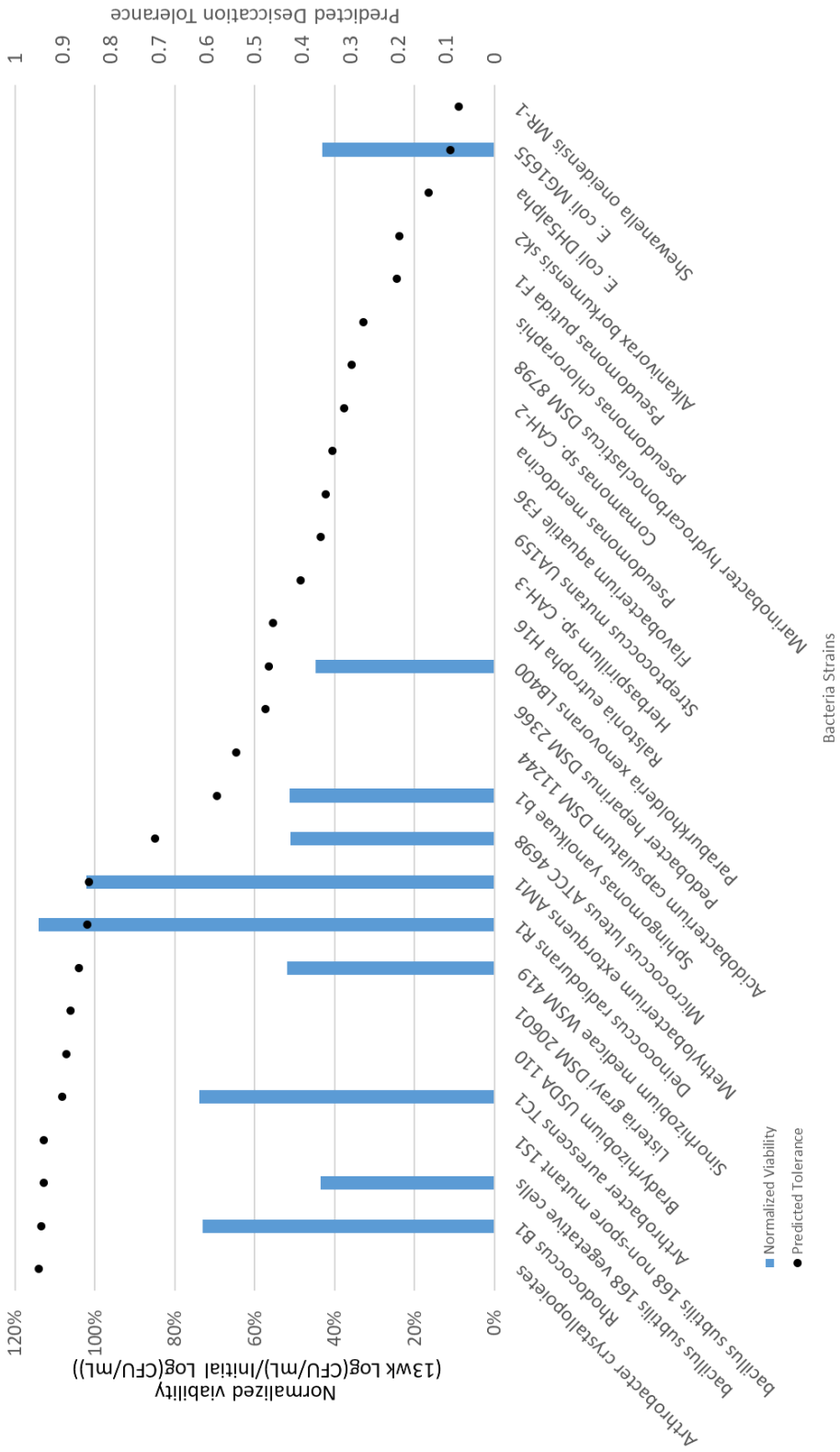

Table A7: Accuracy of the 4 models as verified by the experimental data at 13 weeks. The data was analyzed two ways, first with the results of all 28 species tested in the desiccation assay, second with 24 species. Removed from the analysis were the *B. subtilis* 168 vegetative cells, *B. subtilis* non-spore forming mutant, *B. diazoefficiens* USDA 110, and *S. mutans* UA159. The *B. subtilis* strains were removed because they had originally been used to test the model. The *B. diazoefficiens* was removed because there is an extensive publication history showing its long-term desiccation survival and its lack of survival in this assay was probably an environmental issue. The *S. mutans* UA159 was removed because it is a microaerophile and it is unknown if it was killed by the desiccation or the 20% atmospheric oxygen content.

	Model	Accuracy	95% confidence interval		Sensitivity	Specificity	AUROC	F1-score
All experimental data	Full gene counts	0.75	0.55	0.89	0.72	0.80	0.76	0.70
	Top 30 gene counts	0.71	0.51	0.87	0.67	0.80	0.73	0.67
	Full Binary	0.71	0.51	0.87	0.67	0.80	0.73	0.67
	Top 30 binary	0.75	0.55	0.89	0.67	0.90	0.78	0.72
	Model	Accuracy	95% confidence interval		Sensitivity	Specificity	AUROC	F1-score
Experimental data with selected samples removed	Full gene counts	0.79	0.58	0.93	0.80	0.78	0.79	0.74
	Top 30 gene counts	0.75	0.53	0.90	0.73	0.78	0.76	0.70
	Full Binary	0.75	0.53	0.90	0.73	0.78	0.76	0.70
	Top 30 binary	0.79	0.58	0.93	0.73	0.89	0.81	0.76