PNAS

Supporting Information for

Saccharomycotina yeasts defy longstanding macroecological patterns

Kyle T. David, Marie-Claire Harrison, Dana A. Opulente, Abigail L. LaBella, John F. Wolters, Xiaofan Zhou, Xing-Xing Shen, Marizeth Groenewald, Matt Pennell, Chris Todd Hittinger, & Antonis Rokas

*Antonis Rokas
Email: antonis.rokas@vanderbilt.edu

This PDF file includes:

Figures S1 to S10

Other supporting materials for this manuscript include the following:

Datasets S1 to S10

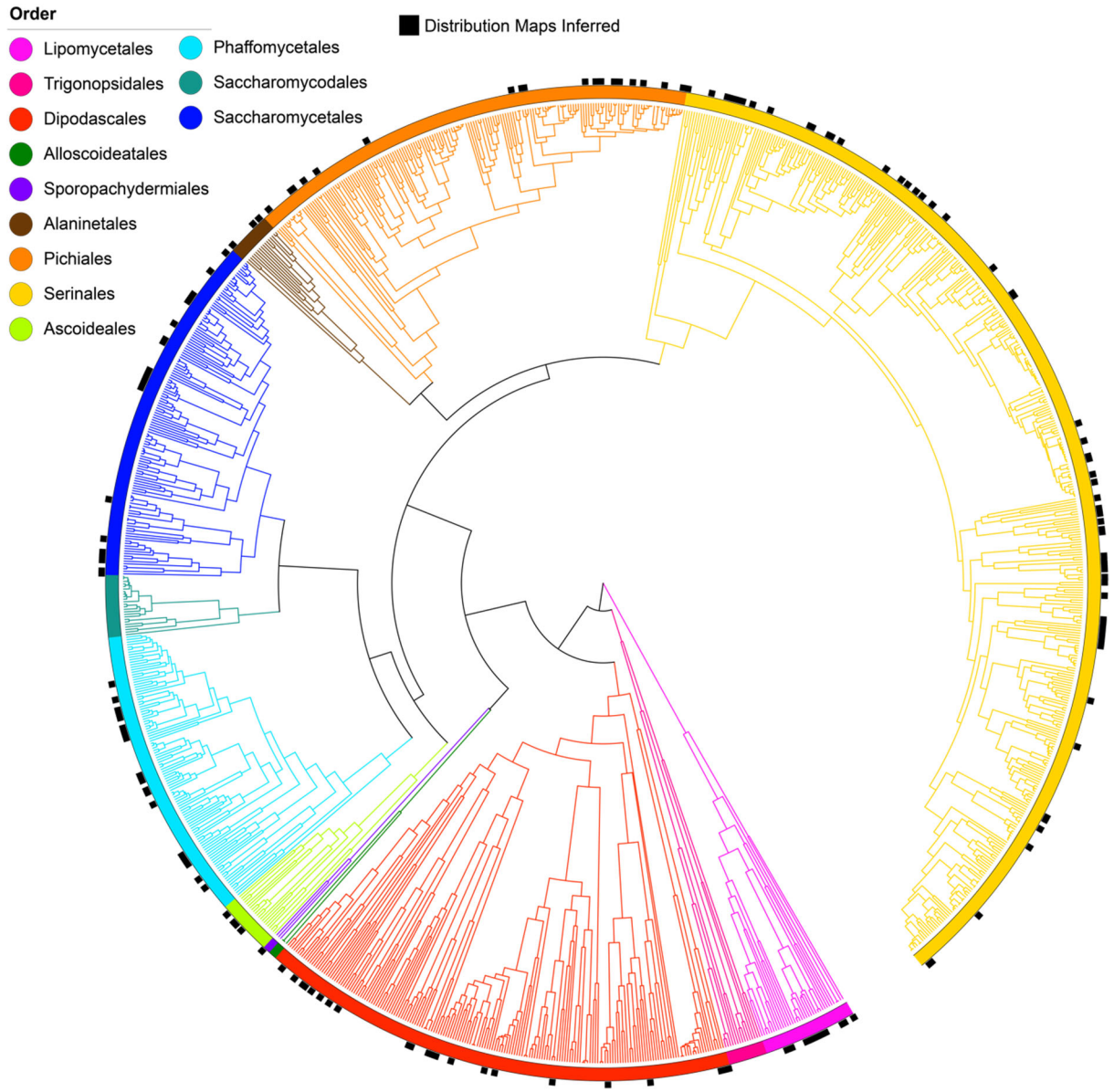

Figure S1. Phylogenetic distribution of sampled species. Species with distribution maps inferred by this study mapped onto a phylogeny highlighting the 12 orders of Saccharomycotina. 11 additional species without phylogenetic information were also included.

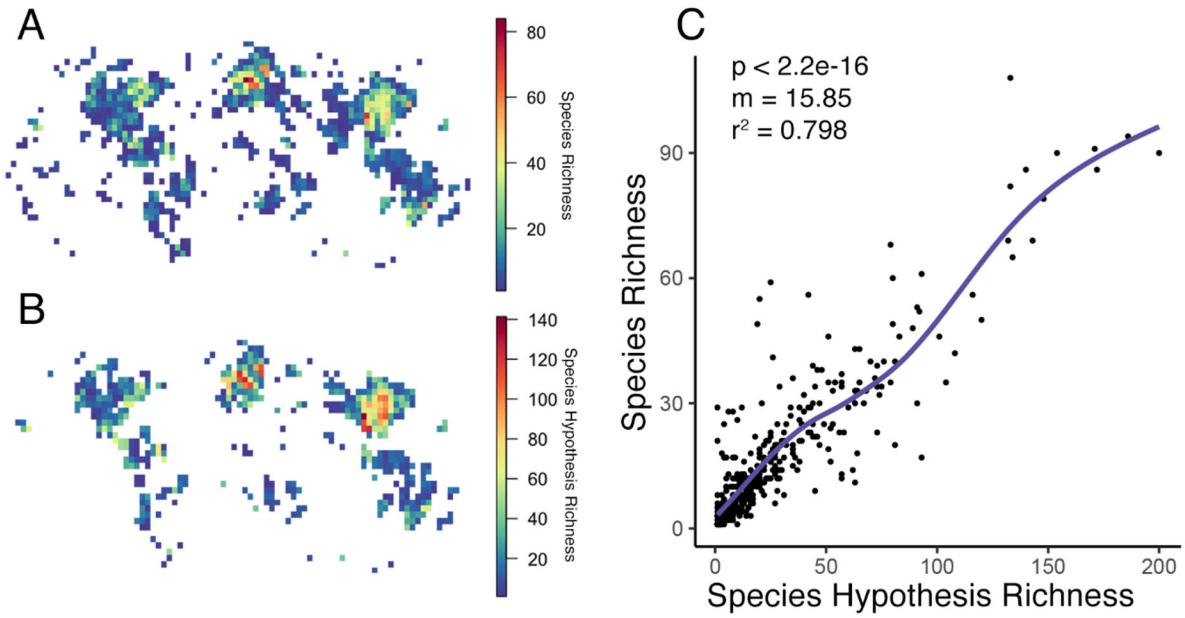

Figure S2. Global patterns of traditional taxonomy and species hypotheses are congruent. A) Geographic heat map of observed species richness for all species considered by this study. B) Geographic heat map of species hypotheses as defined by the UNITE database for molecular identification of fungi. C) Correlation between diversity as estimated by species richness and species hypothesis richness. The p-value, scaled slope, and correlation coefficient of the linear model are displayed.

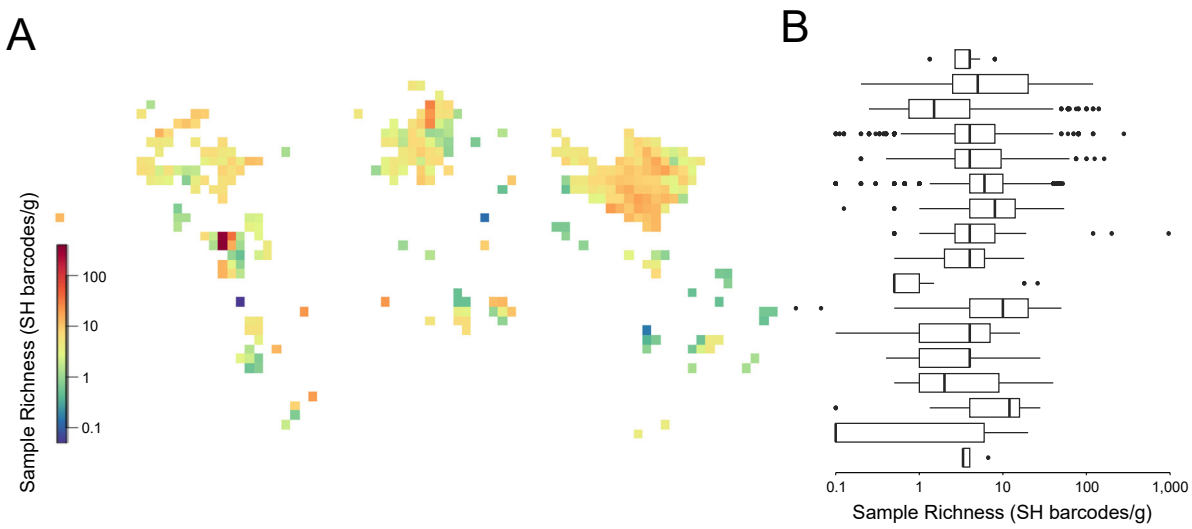

Figure S3. Absence of a latitudinal gradient in soil sample metabarcodes. Species hypothesis richness as defined by the number of unique metabarcodes per gram of sampled soil recover the same major trends as our primary analysis globally (A) and across major latitudinal bands (B). Data provided by GlobalFungi (release 4).

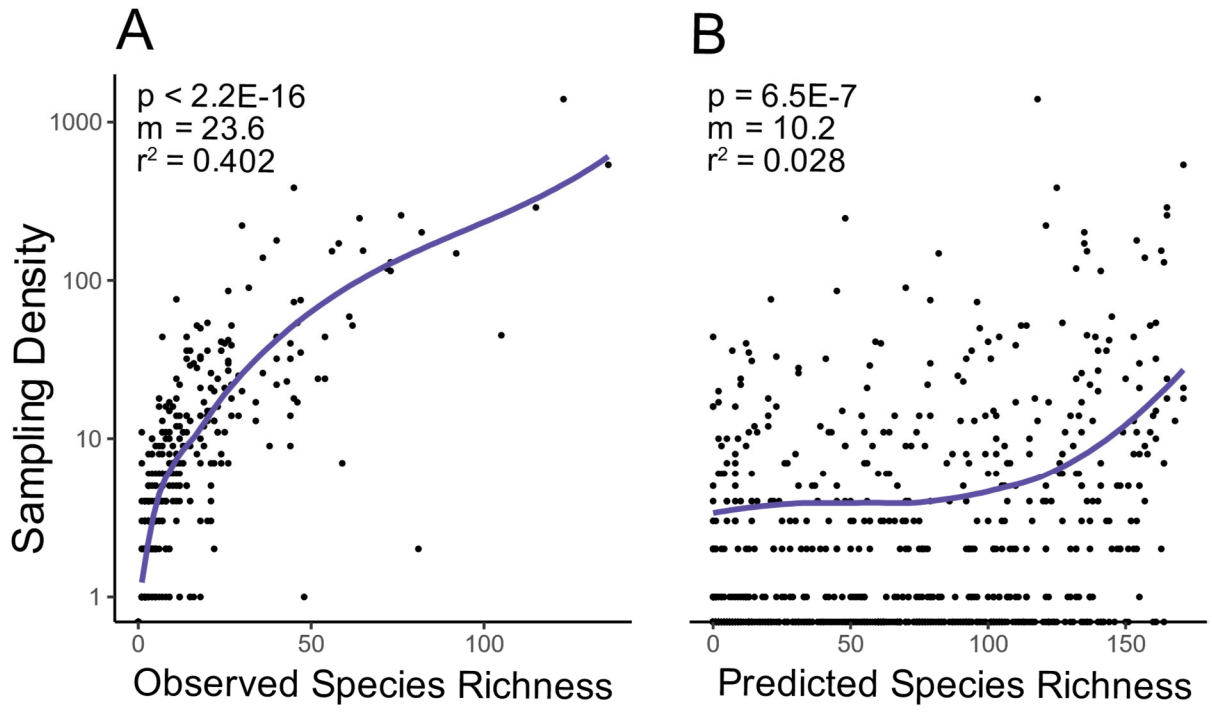

Figure S4. Machine learning reduces sampling bias. Sampling effort of each ecoregion versus A) the number of observed species in the training data and B) the number of predicted species as inferred by random forest species distribution models. The p-value, scaled slope, and correlation coefficient of the linear model are also displayed.

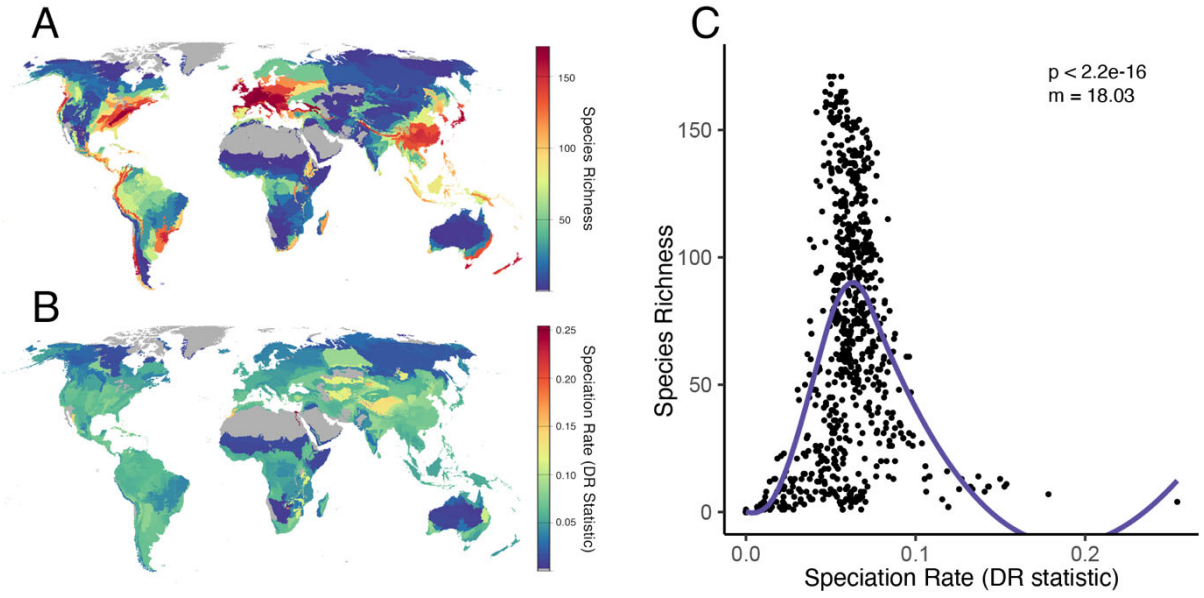

Figure S5. Diversity and diversification are not strongly correlated. A) Species richness per ecoregion. B) Speciation rate per ecoregion, as estimated by the DR species-specific test statistic, weighted by species range. C) Correlation between species richness and speciation rate. The p-value, scaled slope, and correlation coefficient of the linear model are displayed.

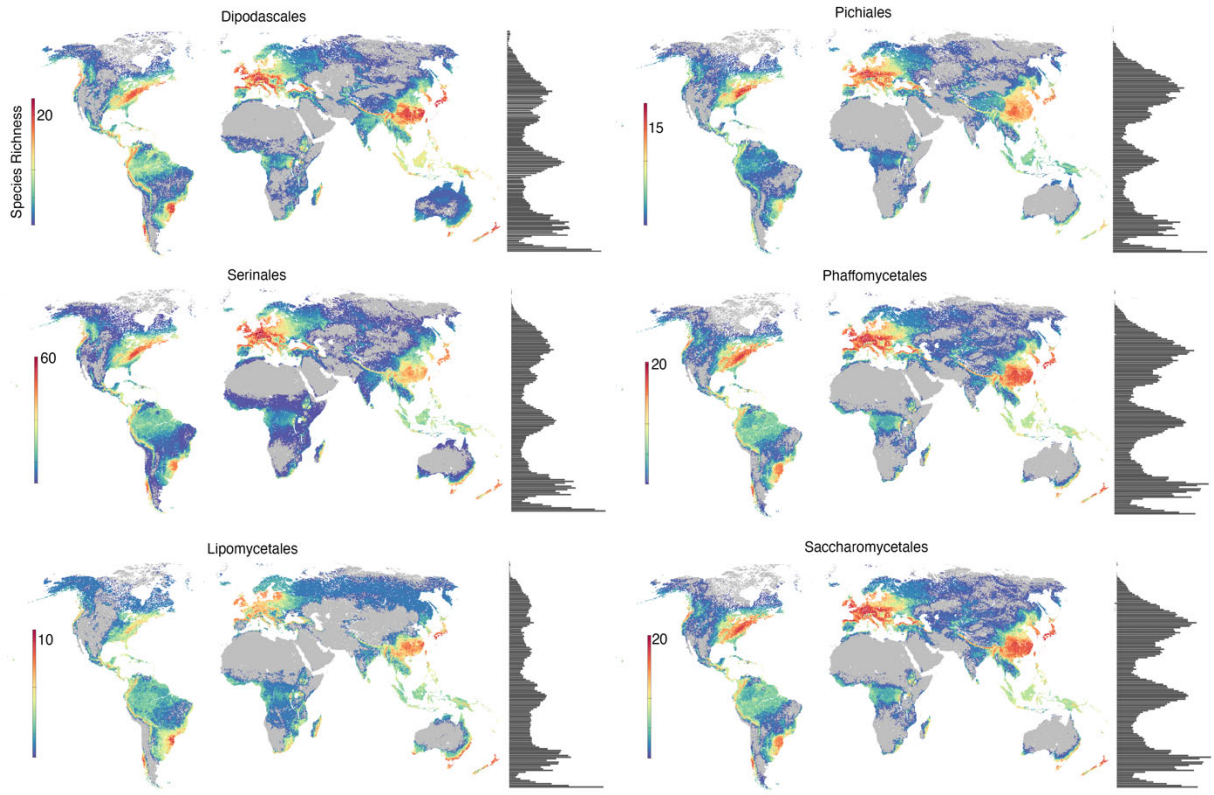

Figure S6. Global diversity patterns are consistent across order. Diversity maps and latitudinal plots for every order in our dataset with at least 10 species.

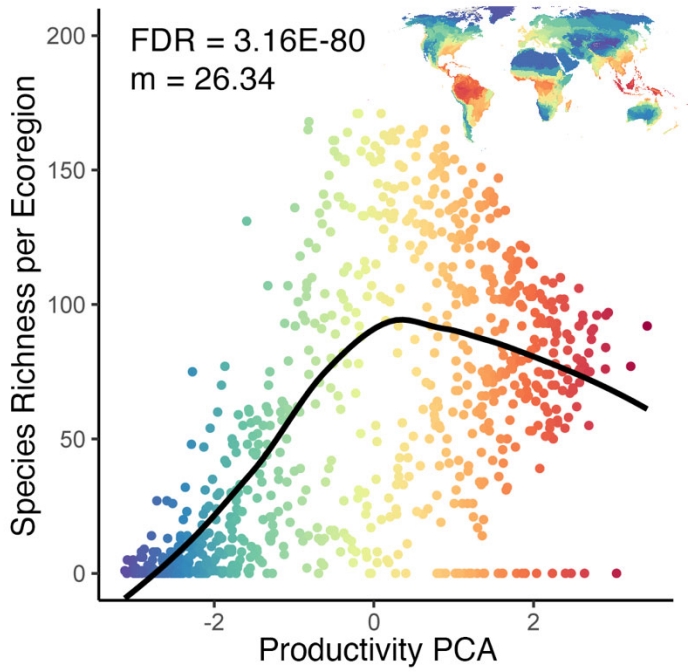

Figure S7. Productivity exhibits a bimodal relationship with species richness.

Colors of points and ecoregions correspond to the x-axis. FDR: false discovery rate of the negative binomial regression. m: scaled slope of linear regression.

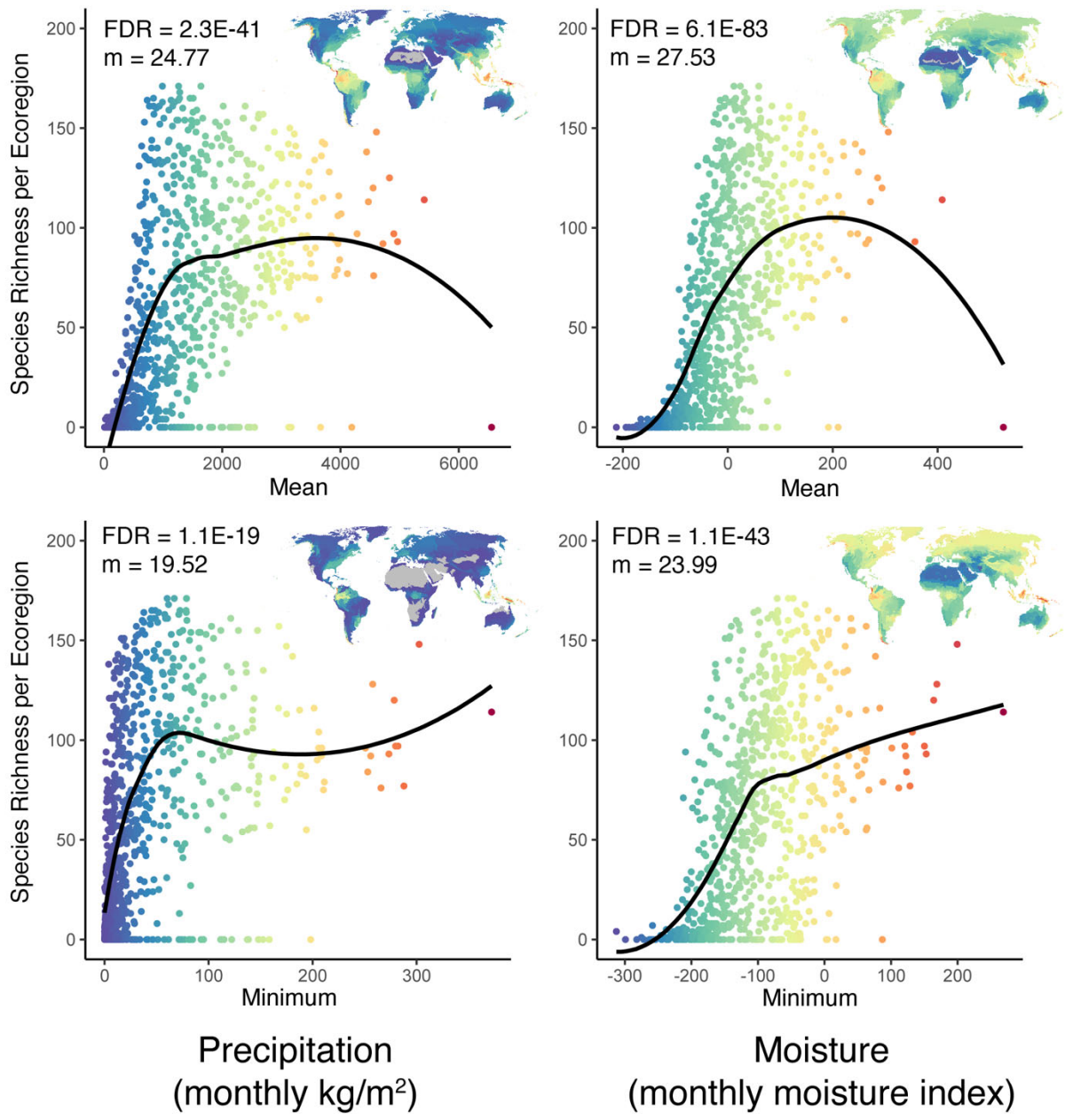

Figure S8. Water content exhibits a logarithmic relationship with area-standardized species richness per ecoregion. Colors of points and ecoregions correspond to the x-axis. FDR: false discovery rate of the negative binomial regression. m: scaled slope of linear regression.

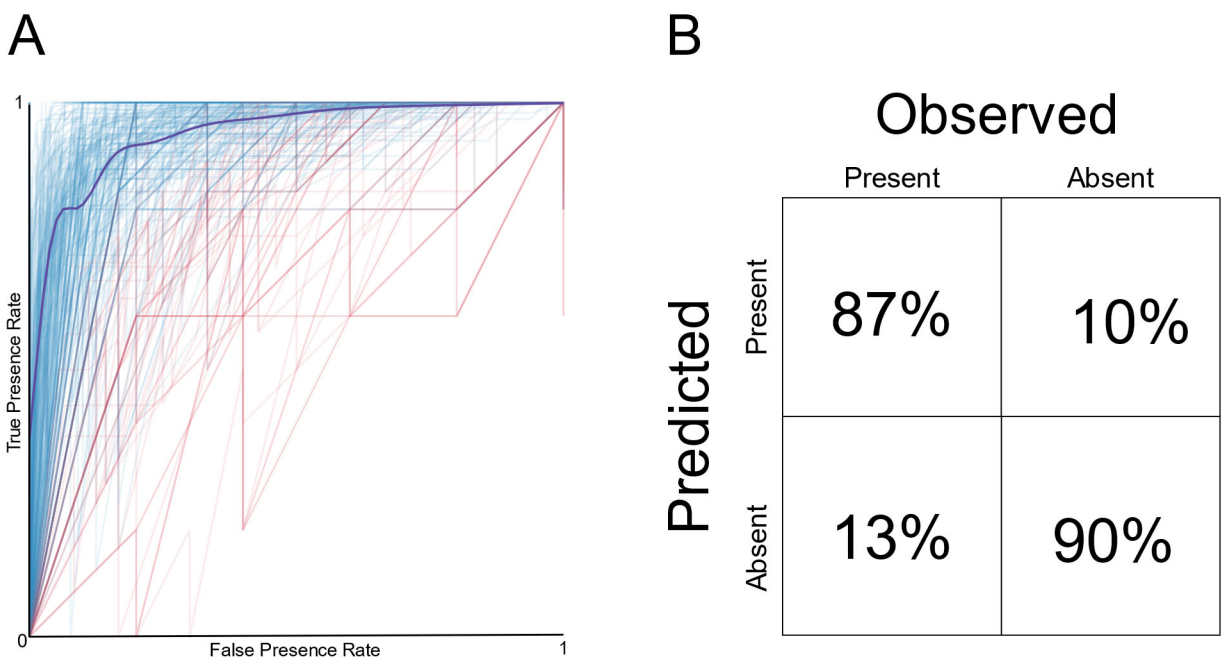

Figure S9. Random forest models accurately identify yeast species distributions. A) Receiver operating characteristic curves of each of the 233 species considered; species in red fell below our threshold and were excluded from analysis. B) Confusion matrix averaged across 186 species included in the analysis.

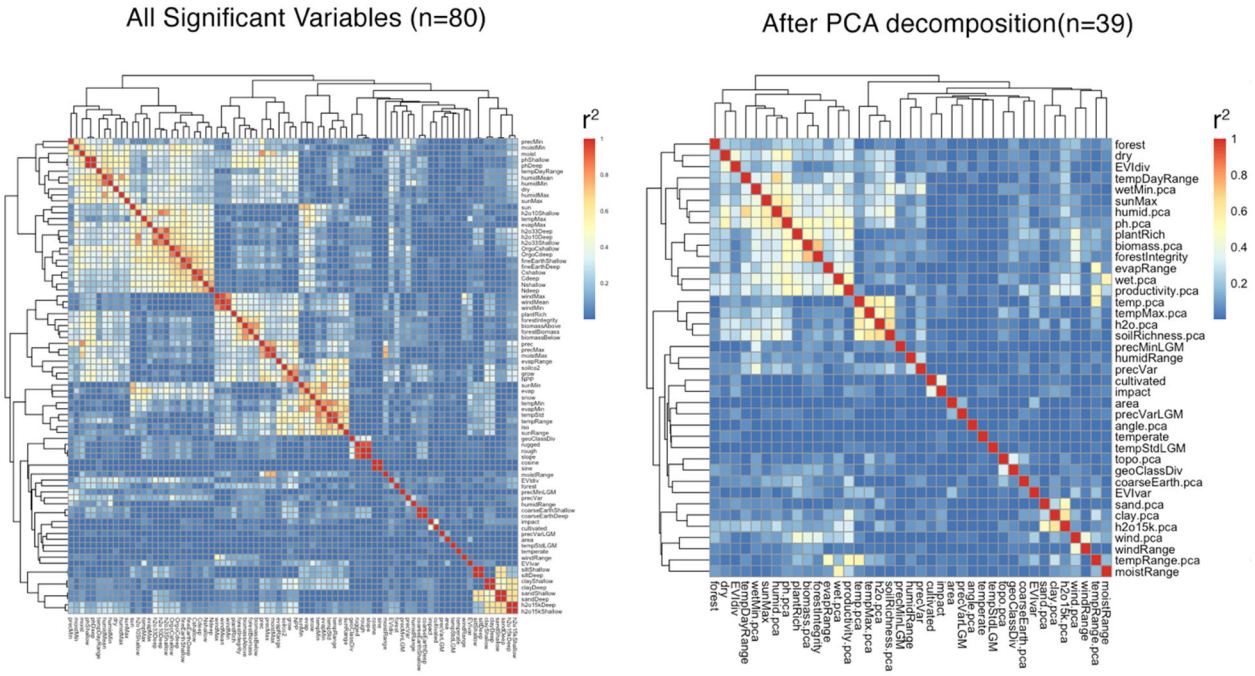

Figure S10. PCA decomposition reduces correlation between environmental predictors. Correlation matrix between each significant environmental variable before and after PCA decomposition.

Dataset S1 (separate file). Summary statistics of soil sample diversity regression. var=predictor name (see Table S6), p=p-value of negative binomial regression, FDR=false discovery rate of negative binomial regression, m=slope of scaled linear regression.

Dataset S2 (separate file). Summary statistics of diversity regression analysis for all variables. var=predictor name (see Table S6), p=p-value of negative binomial regression, FDR=false discovery rate of negative binomial regression, m=slope of scaled linear regression.

Dataset S3 (separate file). Summary statistics of diversity regression analysis for significant variables and principal components. var=predictor name (see Tables S6 and S9), p=p-value of negative binomial regression, FDR=false discovery rate of negative binomial regression, m=slope of scaled linear regression, relative importance=relative importance of top performing predictors.

Dataset S4 (separate file). Relative importance of training variables. Mean decrease in Gini index of every training variable for the 186 species distribution models performed by this study.

Dataset S5 (separate file). Species synonyms. Updated taxonomy used to reconcile previously published occurrence records.

Dataset S6 (separate file). Variable details. Definitions and details for each environmental variable used in training and diversity regression analysis.

Dataset S7 (separate file). Ecoregion data. Aggregated environmental variables for each ecoregion, this data underlies the regression analysis (Datasets S2 and S3).

Dataset S8 (separate file). Binary variable definitions. Definitions for select categorical variables encoded as binomial.

Dataset S9 (separate file). Principal component definitions. Definitions and details for constructed principal components of highly correlated variables.

Dataset S10 (separate file). Range size data. Range size, mean latitude, and mean species overlap for each species, this data underlies the range size phylogenetic comparative modeling analysis (Fig 4).