METHOD FOR IDENTIFYING HIGH ETHANOL TOLERANT YEAST STRAINS BASED ON MULTI OMICS BIOMARKERS

Field of the Invention

The present invention relates to the field of computer technology, particularly to a method for identifying high ethanol tolerant yeast strains based on multi omics biomarkers.

Background to the Invention

Saccharomyces cerevisiae is the core strain for industrial ethanol fermentation. During the ethanol fermentation process, as the ethanol concentration increases, it will have a stress effect on the growth of yeast cells. Yeast cells will produce corresponding stress responses to cope with this stress in order to survive and grow. This coping mechanism is that yeast cells develop tolerant to ethanol. Ethanol tolerant directly determines the fermentation efficiency and product concentration of brewing yeast. However, high concentrations of ethanol can damage cell membrane integrity, inhibit enzyme activity, and trigger oxidative stress, leading to cell growth arrest or even death. It is generally believed that ethanol has three main effects on yeast cells, namely inhibition of cell growth, cell survival, and fermentation. Therefore, ethanol tolerant is often defined by its impact on cell growth. At room temperature, the maximum ethanol concentration allowed for yeast growth after 48 hours of cultivation in a medium containing 1% to 14% ethanol represents the yeast's ethanol tolerant level. Among them, strains that can grow in 3% to 6% ethanol have poor ethanol tolerant, 6% to 10% are moderate, and 10% to 13% are high. This definition method is relatively simple and is commonly used to screen strains with ethanol tolerant. Improving the ethanol tolerant of yeast is of great significance for enhancing fermentation efficiency, yield, and final product quality.

The traditional yeast strain screening method relies on cultivating strains in different ethanol concentration media, and determining the ethanol tolerant of yeast strains by measuring parameters such as growth curve, maximum specific growth rate, and final cell density. However, this method requires a large number of experimental operations, which are time-consuming and labor-intensive, resulting in high screening costs.

Therefore, there is an urgent need for a method to reduce the cost of yeast strain screening.

Statement of Invention

Based on this, it is necessary to provide a method for identifying high ethanol tolerant

yeast strains based on multi omics biomarkers to address the aforementioned technical issues. This method can reduce the screening cost of yeast strains.

The present invention adopts the following technical solution:

The present invention provides a method for identifying high ethanol tolerant yeast strains based on multi omics biomarkers, comprising:

A method for identifying high ethanol tolerant yeast strains based on multi omics biomarkers, comprising:

Obtain gene expression levels, protein expression profiles, and metabolite abundance of multiple unlabeled wild yeast strains;

Perform Z-score normalization on the gene expression level, protein expression profile, and metabolite abundance separately, and align the three sets of data after Z-score normalization according to features;

Extract standardized values of GRE1 gene, TPS1 gene, HSP104 protein, ADH2 protein, CTA1 protein, sphingosine, and trehalose-6-phosphate from three sets of data after feature alignment, and form a multi omics biomarker feature combination based on the extracted standardized values;

Combine the multi omics biomarker features and input them into a strain screening model for industrial strain screening, obtaining a high ethanol tolerant probability score for each of multiple unlabeled wild yeast strains;

Unlabeled wild yeast strains with a high ethanol tolerant probability score greater than or equal to the preset threshold are considered as high ethanol tolerant yeast strains.

Preferably, the multi omics biomarker features are combined and input into a strain screening model for industrial strain screening, obtaining a high ethanol tolerant probability score for each of multiple unlabeled wild yeast strains, specifically comprising:

Arrange the Z-score values of 7 biomarkers in a multi omics biomarker feature combination in order as an input vector;

In the strain screening model, each tree is independently traversed, starting from the root node and splitting according to feature thresholds, unlabeled wild yeast strains are assigned to specific leaf nodes, which correspond to a classification label;

Record the proportion of high ethanol tolerant yeast strains in specific leaf nodes during training as a local probability;

Weighted average the local probabilities of all trees to obtain a weighted average result, which is determined as the high ethanol tolerant probability score of unlabeled wild yeast

strains; the weight of the weighted average is the accuracy of the tree.

Preferably, the training process of the strain screening model specifically includes:

Obtain the sample feature matrix and divide it into a training set and a validation set;

Train the random forest model, support vector machine model, gradient boosting decision tree model, naive Bayes model, neural network model, and generalized linear model using the training set, and adjust the parameters of all models until they are optimal;

For each of the models, input the validation set into the model to obtain the filtering results of the model, compare the filtering results of the model with the true categories of the test set to obtain the indicators of the model; the indicators include prediction accuracy, sensitivity, and specificity;

Compare and analyze the indicators of all models, and determine the optimal model based on the results of the comparative analysis;

Determine the optimal model as the strain screening model.

Preferably, obtaining the sample feature matrix specifically comprises:

Obtain multiple high ethanol tolerant yeast strains and multiple low ethanol tolerant yeast strains as sample training sets;

Obtain the gene expression level, protein expression profile, and metabolite abundance of each yeast strain in the sample training set at different ethanol concentrations;

Perform Z-score normalization on the gene expression level, protein expression profile, and metabolite abundance of the sample to obtain three sets of data;

Perform feature selection on the three sets of data after Z-score normalization to obtain a sample feature matrix; the sample feature matrix includes GRE1 gene, TPS1 gene, HSP104 protein, ADH2 protein, CTA1 protein, sphingosine, and trehalose-6-phosphate.

Preferably, the feature selection is performed on the three sets of data after Z-score standardization processing to obtain a sample feature matrix, specifically comprising:

Standardize the Z-score data and align it across omics using UniProt ID/KEGG ID;

After aligning the cross omics data, perform LASSO regression operation to screen out multiple features;

Based on the selected multiple features, generate an $n \times 7$-dimensional matrix for each yeast strain corresponding to one row and each biomarker corresponding to one column in the feature matrix, and determine the $n \times 7$-dimensional matrix as the sample feature matrix; the n is the number of yeast strains in the sample training set.

Preferably, the screening process of multiple high ethanol tolerant yeast strains and multiple low ethanol tolerant yeast strains specifically includes:

Select candidate strains from the wild-type brewing yeast strain library;

Prepare ethanol containing culture medium;

Cultivate the candidate strain in an ethanol containing medium;

After the cultivation, the candidate strains are initially screened and re screened, and the strains obtained from the re screening are screened according to the screening criteria for high ethanol tolerant strains and low ethanol tolerant strains, resulting in multiple high ethanol tolerant yeast strains and multiple low ethanol tolerant yeast strains.

The present invention provides a device for identifying high ethanol tolerant yeast strains based on multi omics biomarkers, comprising:

Acquisition module, used to obtain gene expression levels, protein expression profiles, and metabolite abundance of multiple unlabeled wild yeast strains;

Alignment module, used for Z-score standardization of gene expression level, protein expression profile, and metabolite abundance, and aligning the three sets of data after Z-score standardization according to features;

The first determination module is used to extract standardized values of GRE1 gene, TPS1 gene, HSP104 protein, ADH2 protein, CTA1 protein, sphingosine, and trehalose-6-phosphate from three sets of data after feature alignment. Based on the extracted standardized values, a multi omics biomarker feature combination is formed;

A screening module is used to combine multiple omics biomarker features and input them into a strain screening model for industrial strain screening, obtaining a high ethanol tolerant probability score for each of multiple unlabeled wild yeast strains;

The second determination module is used to identify unlabeled wild yeast strains with a high ethanol tolerant probability score greater than or equal to a preset threshold as high ethanol tolerant yeast strains.

The present invention provides a computer-readable memory medium that stores a computer program. When the computer program is executed by a processor, the method for determining a high ethanol tolerant yeast strain based on multiple omics biomarkers is implemented.

The present invention provides a computer device comprising a memory, a processor, and a computer program stored on the memory and executable on the processor. When the processor executes the program, it implements the above-mentioned method for

determining high ethanol tolerant yeast strains based on multiple omics biomarkers.

The at least one technical solution adopted in the present invention can achieve the following beneficial effects:

Obtaining gene expression levels, protein expression profiles, and metabolite abundance of multiple unlabeled wild yeast strains, integrating gene expression levels, proteomics, and metabolomics data, providing more comprehensive biological information, and improving the accuracy and reliability of predictions; Perform Z-score normalization on gene expression levels, protein expression profiles, and metabolite abundance, and align the three sets of data after Z-score normalization according to features; Extract standardized values of GRE1 gene, TPS1 gene, HSP104 protein, ADH2 protein, CTA1 protein, sphingosine, and trehalose-6-phosphate from three sets of data after feature alignment, and form a multi omics biomarker feature combination based on the extracted standardized values; Combining multiple omics biomarker features and inputting them into a strain screening model for industrial strain screening, obtaining high ethanol tolerant probability scores for each of multiple unlabeled wild yeast strains, and constructing an efficient strain screening model that overcomes the limitations of traditional methods in data processing and prediction capabilities; Unlabeled wild yeast strains with a high ethanol tolerant probability score greater than or equal to the preset threshold are considered as high ethanol tolerant yeast strains. This method can reduce the screening cost of yeast strains.

Brief Description of the Drawings

The accompanying drawings described herein are intended to provide a further understanding of the present invention and form a part of the present invention. The illustrative embodiments of the present invention and their description are used to explain the present invention and do not constitute undue limitation of the present invention. In the attached figure:

FIG. 1 is a schematic flowchart of a method for identifying high ethanol tolerant yeast strains based on multi omics biomarkers provided by the present invention;

FIG. 2 shows the performance comparison of six machine learning classifiers provided by the present invention;

FIG. 3 is a schematic diagram of the global feature importance provided by the present invention;

FIG. 4 is a schematic diagram comparing the ethanol yield of HT strain and random strain provided by the present invention;

FIG. 5 shows the ROC curve of tolerant prediction for the independent test set (50 strains of bacteria) provided by the present invention;

FIG. 6 is a schematic diagram of the technical route for constructing a multi omics prediction model for ethanol tolerant of brewing yeast provided by the present invention;

FIG. 7 is a schematic diagram of a device for identifying high ethanol tolerant yeast strains based on multi omics biomarkers provided by the present invention;

FIG. 8 is a schematic diagram of a computer device for identifying high ethanol tolerant yeast strains based on multi omics biomarkers provided by the present invention.

Detailed Description

In order to clarify the purpose, technical solution, and advantages of the present invention, the following will provide a clear and complete description of the technical solution of the present invention in conjunction with specific embodiments and corresponding drawings. Obviously, the described embodiments are only a part of the embodiments of the present invention, not all embodiments. Based on the embodiments of the present invention, all other embodiments obtained by ordinary skilled persons in the art without creative labor are within the scope of protection of the present invention.

The traditional yeast strain laboratory screening process involves cultivating yeast strains in ethanol culture media of different concentrations, and determining the ethanol tolerant of yeast strains through parameters such as growth curve, maximum specific growth rate, and final cell density. This method requires a large amount of experimental work, with a long screening cycle, low efficiency, and high cost.

Genetic engineering modification, through gene knockout, gene overexpression, or the use of gene editing techniques such as CRISPR-Cas9, can specifically improve yeast ethanol tolerant. However, this method requires a deep understanding of tolerant related genes, and there are issues with gene stability and transgenic safety. Therefore, analyzing the molecular mechanism of ethanol tolerant and developing accurate prediction methods are of great significance for industrial strain breeding, fermentation process optimization, and synthetic biology modification.

Researchers identified over 250 genes related to ethanol tolerant through whole genome screening, with a focus on vacuoles, peroxisomes, cell walls, and membrane related pathways, and validated the role of FPS1 gene (glycerol channel protein) in ethanol efflux; Researchers combined transcriptome and proteome analysis and found significant changes in yeast mitochondrial and endoplasmic reticulum function, signal transduction (such as GPCR pathway), and aromatic amino acid metabolism under ethanol stress;

Researchers have proposed a differentiation mechanism between high tolerant (HT) and low tolerant (LT) phenotypes through multi omics integration and network analysis, emphasizing the longevity pathway, peroxisome function, energy metabolism, and gene regulation, and establishing a biological mechanism called the "ethanol stress buffering model". Existing research has only analyzed the mechanism of ethanol tolerant through single or multiple omics analyses (genomics, transcriptomics, proteomics, and metabolomics), without involving the construction of predictive models.

Single omics studies, such as using genomics, transcriptomics, proteomics, or metabolomics to investigate the response mechanism of yeast at high ethanol concentrations. Single omics research mainly focuses on understanding how biological information at a single level affects ethanol tolerant. The response mechanism of yeast cells involves multiple levels such as genes, transcription, translation, protein function, metabolism, etc. A single omics data cannot cover all of these levels of information. Single omics research data can only provide partial biological information and cannot fully reflect the response mechanism of yeast at high ethanol concentrations, resulting in limited accuracy and comprehensiveness of prediction results. Multi omics integrated studies, such as genomics, transcriptomics, proteomics, and metabolomics, provide a more comprehensive understanding of the response mechanism of yeast at high ethanol concentrations. Although multi omics integration provides a more comprehensive perspective, these studies mainly focus on mechanism exploration and have not further utilized these data for predictive evaluation.

In addition, traditional methods are limited by experimental conditions, resulting in limited screening efficiency and accuracy of results. In recent years, single omics studies have attempted to elucidate the mechanism of ethanol tolerant, but due to the limited data dimensions, it is difficult to fully reflect complex biological processes, which limits the application of efficient screening.

The present invention aims to solve the above-mentioned technical problems and provide a brewing yeast ethanol tolerant prediction method that integrates gene expression levels, proteomics, metabolomics data, and machine learning algorithms to improve the efficiency and accuracy of strain screening, reduce experimental workload, and guide synthetic biology modification and fermentation process optimization, thereby meeting the needs of industrial production.

The technical solutions provided by various embodiments of the present invention will be described in detail with reference to the accompanying drawings.

Devices such as desktops, servers, laptops, etc. that are capable of implementing the present invention solution. For the sake of convenience, the following explanation will only

focus on the server as the executing entity.

Figure 1 is a schematic flowchart of a method for identifying high ethanol tolerant yeast strains based on multi omics biomarkers in the present invention, which specifically includes the following steps:

S101: Obtain gene expression levels, protein expression profiles, and metabolite abundance of multiple unlabeled wild yeast strains.

S102: Perform Z-score normalization on gene expression levels, protein expression profiles, and metabolite abundance, and align the three sets of data after Z-score normalization according to features.

Cross omics data alignment: Construct a gene protein metabolite module using the Weighted Gene Co expression Network Analysis (WGCNA) package in R language. The gene protein metabolite module refers to a feature set that identifies genes, proteins, and metabolites highly correlated with ethanol concentration through co expression network analysis, reflecting potential regulatory networks or metabolic pathways. Network type: Differentiate signed hybrid regulation, module definition: minimum number of genes=30, cutting height=0.25, screen modules significantly correlated with ethanol concentration ($|r|>0.7$, $p<0.01$).

S103: Extract standardized values of GRE1 gene, TPS1 gene, HSP104 protein, ADH2 protein, CTA1 protein, sphingosine, and trehalose-6-phosphate from three sets of feature aligned data, and form a multi omics biomarker feature combination based on the extracted standardized values.

S104: Combine multiple omics biomarker features and input them into a strain screening model for industrial strain screening, obtaining high ethanol tolerant probability scores for each of multiple unlabeled wild yeast strains.

In an exemplary embodiment, the screening process for multiple high ethanol tolerant yeast strains and multiple low ethanol tolerant yeast strains includes: selecting candidate strains from a wild-type brewing yeast strain library; Prepare ethanol containing culture medium; Cultivate candidate strains in ethanol containing medium; After cultivation, candidate strains are subjected to initial and secondary screening, and the strains obtained from secondary screening are screened according to the screening criteria for high ethanol tolerant strains and low ethanol tolerant strains, resulting in multiple high ethanol tolerant yeast strains and multiple low ethanol tolerant yeast strains.

Specifically, the screening process for multiple high ethanol tolerant yeast strains and multiple low ethanol tolerant yeast strains is as follows:

1、 Source of bacterial strains and pre cultivation

Strain library: Select candidate strains from the wild-type brewing yeast strain library preserved in the laboratory.

Pre cultivation conditions: Resuscitate each strain from glycerol cryovials and inoculate it into YPD liquid medium (glucose 20 g/L, tryptone 20 g/L, yeast extract 10 g/L). Shake at 28 ℃ and 160 rpm for 12 hours until logarithmic growth stage (OD600 ≈ 0.7).

2、 Preparation of ethanol containing culture medium

Culture medium formula: Basic YPD liquid culture medium (same as pre cultivation), sterilized and cooled, and then added with sterile anhydrous ethanol to a final concentration of 13% (v/v).

Control culture medium: YPD medium without ethanol, used to evaluate the basic growth performance of the strain.

3、Initial screening: Solid tablet tolerant test

Gradient dilution: Dilute the pre cultured bacterial solution in a gradient of 10 ^ -3~10 ^ -5, take 100 μL and apply it to YPD solid plates containing 13% ethanol and ethanol free control plates.

Cultivation and observation: Incubate at 28 ℃ for 48-72 hours, screening criteria:

Initial screening of HT candidate strains: Clear colonies (colony diameter ≥ 1 mm) were formed on a 13% ethanol plate, and the difference in colony count compared to the control group was ≤ 50%.

Initial screening of LT candidate strains: There were no visible colonies on the 13% ethanol plate or the number of colonies decreased by ≥ 90% compared to the control group.

4、 Re screening: Analysis of liquid culture growth curve

Inoculation and cultivation: Inoculate the HT and LT candidate strains obtained from the initial screening into YPD liquid medium containing 13% ethanol at a 3% inoculation amount (OD600 ≈ 0.7), while setting up an ethanol free control group.

OD600 dynamic monitoring: Shake the culture at 28 ℃ and 160 rpm, measure the OD600 value every 2 hours until 24 hours, and plot the growth curve.

5、 Screening criteria:

HT strain: Delay period ≤ 6 hours in 13% ethanol, logarithmic growth period OD600

growth rate ≥ 0.1/h, final OD600 value ≥ 60% of the control group.

LT strain: Delay period ≥ 12 hours, OD600 growth rate ≤ 0.05/h, final OD600 value ≤ 20% of the control group.

6、Tolerant verification

(1) Cell survival rate determination:

Cultivate HT and LT candidate strains in 13% ethanol medium for 24 hours, and then coat them on ethanol free YPD plates after gradient dilution.

Calculate survival rate: Survival rate (%)=(CFU of ethanol treatment group/CFU of control group) × 100.

HT criteria: Survival rate ≥ 50%; LT standard: Survival rate ≤ 10%.

2. Observation of cell morphology:

Take the bacterial solution treated with 13% ethanol for 24 hours, collect the bacterial cells by centrifugation, fix them with 2.5% glutaraldehyde, and observe them under scanning electron microscopy.

HT strain: complete cell morphology (elliptical, smooth surface); LT strain: Cell rupture or severe wrinkling.

Data collection and training set construction:

Select 10 strains of bacteria (biological replicates) from the HT/LT group and 5 ethanol concentrations as training set samples.

a、Gene sequencing: RNA-Seq technology is used to perform high-throughput sequencing of gene expression mRNA in brewing yeast cells at different ethanol concentrations, screening for genes that may be related to ethanol tolerant.

(1) Experimental design:

Collect mid logarithmic growth stage cells (OD600 ≈ 0.6) from the training set strains and rapidly fix them with RNA protectants (such as RNAlater). Preparation using chain specific libraries (retaining chain direction information). Use Illumina NovaSeq 6000 and PE150 mode for sequencing, requiring a data volume of ≥ 20M clean reads/sample (to ensure detection of low abundance genes).

(2) Bioinformatics analysis:

Use FastP software to perform quality control on the raw sequencing data and filter out low-quality sequences and sequencing adapter sequences. Using hisat2 software,

Saccharomyces cerevisiae S288C (NCBI RefSeq ID: GCF000146045.2) was used as the reference genome for sequence alignment. Transcripts were assembled and quantified using stringgie software, and TPM quantification of each gene was output.

b、 Quantitative proteomic analysis: Analyze the protein expression profile of brewing yeast under different ethanol concentrations using liquid chromatography-mass spectrometry (LC-MS).

(1) Experimental design:

In quantitative proteomics analysis, brewing yeast cells were lysed, reduced alkylated, hydrolyzed with Trypsin Gold enzyme, labeled with TMT 16plex peptide segments, graded by high pH reverse phase chromatography, and detected in DDA mode using a Q Exactive HF-X mass spectrometer (MS1 120000 resolution, MS2 45, 000 resolution).

(2) Data analysis:

The raw data was processed using the Andromeda search engine with MaxQuant (v2.1.0) software. The parameters were set as follows: parent ion mass tolerant of 4.5ppm, fragment ion tolerant of 20ppm, enzyme digestion method of Trypsin/P (allowing up to 2 cleavage sites), fixed modifications of TMT 16plex (N-terminal and lysine) and cysteine alkylation (+57.021Da), variable modifications including methionine oxidation (+15.995Da) and protein N-terminal acetylation (+42.011Da). The UniProt brewing yeast reference protein library (May 2024 version, containing common pollutants) was selected as the database, and FDR<1% was calculated using the reverse bait library as the identification threshold.

c、 Quantitative metabolomics analysis: Obtain the abundance changes of metabolites in fermentation broth under different ethanol concentrations through mass spectrometry analysis.

(1) Experimental Design

The sample pretreatment used pre cooled methanol acetonitrile water (4:4:2, v/v/v) to quench metabolic activity, ultrasound assisted extraction (300W, 2s/3s pulse, 5min), centrifugation (15000g, 15min, 4 ℃), and nitrogen blow concentration of the supernatant. Adopting an ultra-high performance liquid chromatography (UPLC) system, paired with two types of chromatography columns:

Polar metabolite: ACQUITY UPLC BEH Amide column (2.1 × 100mm, 1.7 μ m), mobile phase is 25mM ammonium acetate+ammonia solution (pH 9.0)/acetonitrile.

Non polar metabolites: ACQUITY UPLC HSS T3 column (2.1 × 100mm, 1.7 μ m), mobile phase 0.1% formic acid water/acetonitrile mass spectrometry detection uses Q-TOF

6600+system, ESI positive and negative ion mode switching scanning, parameter settings: ion source temperature 500 ℃, spray voltage ± 5.5kV, collision energy 20-40V, mass range m/z 50-1500, resolution>30000. Add an internal standard (L-2-chlorophenylalanine) to monitor data stability in each batch of experiments.

The raw mass spectrometry data was subjected to peak extraction, alignment, and normalization (internal standard correction) using Progenesis QI. Metabolite identification is achieved through precise mass number (error<5ppm) and secondary spectrum matching (METLIN/HMDB database, similarity>80%).

Data preprocessing:

Standardization processing: Due to significant dimensional differences in different omics data (such as gene TPM values ranging from 0 to 10 ^ 5 and metabolite peak areas ranging from 10 ^ 3 to 10 ^ 8), this invention uses R language to perform Z-score standardization on gene expression levels, protein and metabolite abundance, eliminating biases in different omics quantitative methods and experimental batches.

Dimensionality reduction: Apply principal component analysis (PCA) dimensionality reduction techniques to reduce data dimensions and preserve key information.

In an exemplary embodiment, the training process of the strain screening model specifically includes: obtaining a sample feature matrix and dividing the sample feature matrix into a training set and a validation set; Train the random forest model, support vector machine model, gradient boosting decision tree model, naive Bayes model, neural network model, and generalized linear model using the training set, and adjust the parameters of all models until they are optimal; For each model, input the validation set into the model, obtain the filtering results of the model, compare the filtering results of the model with the true categories of the test set, and obtain the indicators of the model; The indicators include prediction accuracy, sensitivity, and specificity; Compare and analyze the indicators of all models, and determine the optimal model based on the results of the comparative analysis; Determine the optimal model as the strain screening model.

In an exemplary embodiment, obtaining a sample feature matrix specifically includes: obtaining multiple high ethanol tolerant yeast strains and multiple low ethanol tolerant yeast strains as sample training sets; Obtain the gene expression level, protein expression profile, and metabolite abundance of each yeast strain in the sample training set at different ethanol concentrations; Z-score normalization was performed on the gene expression level, protein expression profile, and metabolite abundance of the samples to obtain three sets of data; Perform feature selection on the three sets of data after Z-score standardization to obtain a sample feature matrix; The sample feature matrix includes GRE1 gene, TPS1 gene, HSP104 protein, ADH2 protein, CTA1 protein, sphingosine, and

trehalose-6-phosphate.

Specifically, the caret package based on R language of the present invention adopts six machine learning algorithms for parallel training to comprehensively evaluate model performance. The six machine learning algorithms are as follows:

Random Forest (RF): Multiple decision trees are constructed based on Bootstrap sampling, and classification results are output through a voting mechanism. The default number of trees is set to 500, making it suitable for high-dimensional features and resistant to overfitting.

Support Vector Machine (SVM): Select radial basis kernel function to achieve nonlinear classification by maximizing the classification interval, and adjust the penalty parameter C and kernel parameter γ.

Gradient Boosting Decision Tree (GBDT): optimizes decision tree residuals iteratively, with a default learning rate of 0.1 and a tree depth of 3. It is suitable for scenarios with complex interactions between features.

Naive Bayes Model (NBM): Based on Bayes' theorem, assuming independent features, suitable for small sample data, with a probability threshold of 0.5.

Neural Network (NN): Construct a single hidden layer fully connected network, with 1.5 times the number of hidden layer nodes as the number of features, ReLU as the activation function, and Adam as the optimizer.

Generalized Linear Model (GLM): Using logistic regression (binary classification), L2 regularization to control coefficient weights, and a link function of logit.

The input data for the six models needs to be standardized (Z-score), with category labels (HT/LT) encoded as 1/0, and the training and testing sets divided into layers in a 7:3 ratio.

Optimize six models.

The pROC package based on R language optimizes the hyperparameters of the model through ten fold cross validation, reducing overfitting and improving the model's generalization ability.

1. The range of hyperparameter tuning is:

RF: Adjust mtry (number of randomly selected features per tree, range 1-7) and ntree (100-500).

SVM: Grid search for C (0.1-10) and γ (0.01-0.1).

GBDT: Optimize nrounds (50-200 iterations), max_depth (3-6), and eta (learning rate,

0.01-0.3).

NN: Adjust the number of hidden layer nodes (5-15) and dropout rate (0.2-0.5).

GLM: Regularization parameter lambda (0.001-1).

2. The cross validation process is as follows:

Divide the training set into 10 equal parts, alternating between 9 training parts and 1 validation part, and repeat 10 times.

Record the AUC, sensitivity (recall), and specificity of each fold, and take the average as the basis for hyperparameter selection.

The early stop strategy prevents overfitting, for example, terminating when the GBDT iteration loss no longer decreases.

3. Performance comparison and model selection:

Figure 2 shows the performance comparison of six machine learning classifiers provided by the present invention. As shown in Figure 2, the RF model is significantly better than other models in terms of ROC, sensitivity, and specificity when comprehensively comparing the six models from three dimensions: prediction accuracy, sensitivity, and specificity, $p<0.05$, Paired t-test with high training efficiency, less than 5 minutes.

The core hyperparameters of the random forest are shown in Table 1:

Table 1

Hyper-parameters	Numerical value	Function explanation
n_estimators	500	Number of decision trees
max_depth	5	Maximum depth of a single tree, controlling overfitting
min_samples_split	5	Minimum sample size required for node splitting
min_samples_leaf	5	Minimum sample size for leaf nodes to prevent noise interference
max_features	“sqrt”	The maximum number of features when each tree splits
bootstrap	True	Is there a use of replacement sampling to maintain diversity
class_weight	“balanced”	Dealing with HT/LT category imbalance
random_state	42	Fixed random number seed to ensure reproducibility of results

In an exemplary embodiment, feature selection is performed on three sets of data after Z-score normalization to obtain a sample feature matrix, which specifically includes aligning the Z-score standardized data across omics using UniProt ID/KEGG ID; After cross omics data alignment, perform LASSO regression operation to screen out multiple features; Based on the selected multiple features, generate an $n \times 7$-dimensional matrix for each yeast strain corresponding to one row and each biomarker corresponding to one column in the feature matrix, and determine the $n \times 7$-dimensional matrix as the sample feature matrix; n is the number of yeast strains in the sample training set.

Specifically, the feature selection process is as follows:

Cross omics data alignment: Construct a gene protein metabolite module using the Weighted Gene Co expression Network Analysis (WGCNA) package in R language. The gene protein metabolite module refers to a feature set that identifies genes, proteins, and metabolites highly correlated with ethanol concentration through co expression network analysis, reflecting potential regulatory networks or metabolic pathways. Network type: Differentiate signed hybrid regulation, module definition: minimum number of genes=30, cutting height=0.25, screen modules significantly correlated with ethanol concentration ($|r|>0.7$, $p<0.01$).

LASSO regression: Using LASSO regression in the R language glmnet package to further screen out the features that contribute the most to the prediction model, reducing the risk of overfitting.

The present invention integrates the following seven multi omics biomarker feature combinations:

Genes: GRE1 gene, TPS1 gene;

Proteins: Heat shock protein 104 (HSP104), alcohol dehydrogenase 2 (ADH2), catalase 1 (CTA1);

Metabolites: Sphingosine, Trehalose-6-phosphate.

Tag: Classify HT/LT as a binary tag.

In an exemplary embodiment, multiple omics biomarker features are combined and input into a strain screening model for industrial strain screening, obtaining a high ethanol tolerant probability score for each of multiple unlabeled wild yeast strains. This includes: arranging the Z-score values of seven biomarkers in the multi omics biomarker feature combination in order as an input vector; In the strain screening model, each tree is independently traversed, starting from the root node and splitting according to feature thresholds. Unlabeled wild yeast strains are assigned to specific leaf nodes, which

correspond to a classification label; Record the proportion of high ethanol tolerant yeast strains in specific leaf nodes during training as a local probability; Weighted average the local probabilities of all trees to obtain a weighted average result, which is determined as the high ethanol tolerant probability score of unlabeled wild yeast strains; The weight of the weighted average is the accuracy of the tree.

Specifically, the SHAP values of standardized multi omics feature matrices (7 key biomarkers: 2 genes, 3 proteins, 2 metabolites) were calculated using the trained optimal random forest model (RF, AUC=0.95) using the R language SHAPR package to evaluate the importance of features to the model. A SHAP importance map was drawn, and Figure 3 shows the global feature importance diagram provided by the present invention. The bar graph displays the average | SHAP | value of features, and the results show that GRE1 gene, CTA1, and HSP104 are the top 3 contributing features.

Specifically, probability scoring is used to guide the screening of industrial strains, which can identify candidate strains suitable for high ethanol fermentation.

S105: Unlabeled wild yeast strains corresponding to high ethanol tolerant probability scores greater than or equal to a preset threshold are identified as high ethanol tolerant yeast strains.

Specifically, the preset threshold is set according to specific engineering practices. In an industrial application example, the trained random forest model was applied to multi omics data of 200 unlabeled wild yeast strains (including gene sequencing, proteomics, and metabolomics data), and the 'predict()' function was used to output the HT probability score (threshold ≥ 0.8) for each strain. Finally, 19 high probability HT candidate strains were selected; After small-scale fermentation verification (10% ethanol stress, 48 hours of cultivation at 30 ℃); Figure 4 is a schematic diagram comparing the ethanol yield of HT strains and random strains provided by the present invention. As shown in Figure 4, the ethanol yield of these strains averaged 11.8 ± 0.8 g/L, which was significantly increased by 18% compared to the randomly selected control group (10.0 ± 1.2 g/L) (independent sample t-test, $p<0.01$）， And the expression level of key tolerant genes (such as HSP104) was upregulated by 2.1-3.5 times, confirming that model guided targeted screening can effectively improve the performance of industrial strains.

In an exemplary embodiment, in independent test set validation, 50 pre reserved multi omics data of known HT/LT phenotype brewing yeast strains (including 7 biomarker features: GRE1 gene, HSP104, etc.) were used as inputs, and a trained random forest model was loaded for prediction. The ROC curve of the predicted probability and the true label was calculated using the roc() function of the pROC package. The results showed that the AUC reached 0.88 (95% CI: 0.82-0.93), the sensitivity was 0.85, and the

specificity was 0.83. Figure 5 was plotted using ggplot2, Figure 5 is a schematic diagram of the ROC curve for predicting the tolerant of an independent test set (50 strains of bacteria) provided by the present invention. As shown in Figure 5, the ROC curve has false positive rate on the horizontal axis, true positive rate on the vertical axis, and diagonal line as a random reference. The area under the curve is significantly higher than random guess (DeLong test p=2.3 × 10 ^ -6), confirming that the model has stable predictive ability for ethanol tolerant of unknown strains. Subsequently, decision curve analysis (DCA) was conducted to demonstrate its clinical application value in industrial strain screening.

In an exemplary embodiment, the present invention provides a technical roadmap for constructing a multi omics prediction model for ethanol tolerant of brewing yeast as shown in Figure 6. As shown in Figure 6, the multi omics prediction model for ethanol tolerant of brewing yeast includes a data acquisition stage, a feature engineering stage, and a modeling application stage. In the data collection stage, bacterial strain screening and cultivation, as well as multi omics data collection, are carried out to obtain sample label data and multi omics quantitative data; In the feature engineering stage, multiple omics data standardization and feature selection are performed to obtain standardized feature matrices and key biomarkers; In the modeling application stage, construct feature matrices and machine learning models, determine the optimal model as RF, and determine the SHAP value representing feature importance; Screening of wild strains and detection of ethanol yield based on the optimal model and SHAP value.

When applying the method for identifying high ethanol tolerant yeast strains based on multi omics biomarkers provided by the present invention, the order of each step shown in Figure 1 may not be followed. The specific order of each step can be determined as needed, and the present invention does not limit this.

Given the current state of technology, the advantages of the present invention are as follows:

Multi level data integration: This invention integrates genetic, proteomic, and metabolomic data, providing more comprehensive biological information and improving the accuracy and reliability of predictions.

Application of machine learning: By applying machine learning algorithms, the present invention can extract key features from multiple omics data, construct efficient prediction models, and overcome the limitations of traditional methods in data processing and prediction capabilities.

Industrial application orientation: This invention focuses on providing a predictive tool that can be directly applied to industrial strain screening, fermentation process optimization,

and synthetic biology modification, outputting tolerant scores that can be directly used in production. Improved the practicality and operability of the technology.

Improving screening efficiency and reducing costs: By using machine learning models for prediction, experimental workload and time costs are reduced, and the efficiency of strain screening and modification is improved.

The above is a method for identifying high ethanol tolerant yeast strains based on multi omics biomarkers provided by one or more embodiments of the present invention. Based on the same idea, the present invention also provides a corresponding device for determining a high ethanol tolerant yeast strain based on multiple omics biomarkers, as shown in Figure 9.

Figure 7 is a schematic diagram of a device for identifying high ethanol tolerant yeast strains based on multi omics biomarkers provided by the present invention, comprising:

Acquisition module 701, used to obtain gene expression levels, protein expression profiles, and metabolite abundance of multiple unlabeled wild yeast strains;

Alignment module 702, used for Z-score standardization of gene expression level, protein expression profile, and metabolite abundance, and aligning the three sets of data after Z-score standardization according to features;

The first determination module 703 is used to extract standardized values of GRE1 gene, TPS1 gene, HSP104 protein, ADH2 protein, CTA1 protein, sphingosine, and trehalose-6-phosphate from three sets of data after feature alignment. Based on the extracted standardized values, a multi omics biomarker feature combination is formed;

A screening module 704 is used to combine multiple omics biomarker features and input them into a strain screening model for industrial strain screening, obtaining a high ethanol tolerant probability score for each of multiple unlabeled wild yeast strains;

The second determination module 705 is used to identify unlabeled wild yeast strains with a high ethanol tolerant probability score greater than or equal to a preset threshold as high ethanol tolerant yeast strains.

The specific limitations of a device for identifying high ethanol tolerant yeast strains based on multi omics biomarkers can be found in the previous section on the limitations of a method for identifying high ethanol tolerant yeast strains based on multi omics biomarkers, which will not be repeated here. The various modules in the device for determining high ethanol tolerant yeast strains based on multiple omics biomarkers can be fully or partially implemented through software, hardware, and their combinations. The above modules can be embedded in hardware form or independent of the processor in

the computer device, or stored in software form in the memory of the computer device, so that the processor can call and execute the corresponding operations of the above modules.

The present invention also provides a computer-readable memory medium storing a computer program that can be used to execute the method for determining a high ethanol tolerant yeast strain based on multiple omics biomarkers provided in Figure 3.

The present invention also provides a schematic diagram of the structure of the computer device shown in Figure 8. As shown in Figure 8, at the hardware level, the computer device includes a processor, internal bus, network interface, memory, and non-volatile memory, and may also include hardware required for other services. The processor reads the corresponding computer program from non-volatile memory into the memory and runs it to implement the method for identifying high ethanol tolerant yeast strains based on multi omics biomarkers provided in Figure 1 above.

Ordinary technical personnel in this field can understand that implementing all or part of the processes in the above embodiments can be accomplished by instructing relevant hardware through a computer program. The computer program can be stored in a non-volatile computer-readable memory medium, and when executed, it may include the processes of the embodiments of the above methods. Among them, any reference to memory, memory, database or other media used in the embodiments provided by the present invention may include at least one of non-volatile and volatile memory. Non volatile memory can include Read Only Memory (ROM), magnetic tape, floppy disk, flash memory, or optical memory. Volatile Memory can include Random Access Memory (RAM) or external cache memory. As an illustration and not a limitation, RAM can take various forms, such as Static Random Access Memory (SRAM) or Dynamic Random Access Memory (DRAM).

The various technical features of the above embodiments can be combined arbitrarily. In order to make the description concise, not all possible combinations of the various technical features in the above embodiments have been described. However, as long as there is no contradiction in the combination of these technical features, they should be considered within the scope of the present invention.