JOURNAL OF AGRICULTURAL AND FOOD CHEMISTRY

pubs.acs.org/JAFC

Article

Using Machine Learning to Identify Biomarkers Affecting Fat Deposition in Pigs by Integrating Multisource Transcriptome Information

Huatao Liu, Kai Xing, Yifan Jiang, Yibing Liu, Chuduan Wang,* and Xiangdong Ding*

Cite This: *J. Agric. Food Chem.* 2022, 70, 10359−10370

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Fat deposition in pigs is not only closely related to pig production efficiency and pork quality but also an ideal model for human obesity. Transcriptome sequencing is widely used to study fat deposition. However, due to small sample sizes, high false positive rates, and poor consistency of results from different studies, new strategies are urgently needed. Machine learning, a new analysis method, can effectively fit complex data and accurately identify samples and genes. In this study, 36 samples of adipose tissue, muscle tissue, and liver tissue were collected from Songliao black pigs and Landrace pigs, and the mRNA of all the samples was sequenced. In addition, we collected transcriptome data for 64 samples in the GEO database from four different sources. After standardization and imputation of missing values in the data set comprising 100 samples, traditional differential expression analysis was carried out, and different numbers of expressed genes were selected as features for the training model of eight machine learning methods. In the 1000 replications of fourfold cross validation with 100 samples, AdaBoost performed best, with an average prediction accuracy greater than 93% and the highest mean area under the curve in predicting the high- and low-fat content groups among the eight ML methods. According to their performance-based ranks inferred by AdaBoost, 12 genes related to fat deposition were identified; among them, *FASN* and *APOD* were specifically expressed in adipose tissue, and *APOA1* was specifically expressed in the liver, which could be important candidate biomarkers affecting fat deposition.

KEYWORDS: *fat deposition, pigs, data integration, machine learning, biomarkers*

INTRODUCTION

With the improvement of living standards, people pay more and more attention to the quality of meat. As the main source of meat, pork is closely related to human health, and fat deposition in pigs is closely related to pork quality and yield.[1] Therefore, it has become an important topic among scientists to improve the meat quality and yield of pigs by exploring the mechanism of fat deposition. Moreover, due to their physiological similarity with humans, pigs have gradually become an ideal model animal for the study of human obesity and metabolic syndrome.[2] Fat deposition is a dynamic equilibrium process involving the synthesis, breakdown, and transport of fat that takes place mainly in adipose tissue, liver, and muscle.[3] In addition, fat deposition is temporally and spatially regulated by multiple genes. Transcriptome sequencing data from different tissues at different times have been widely used to explore the mechanisms of fat deposition. However, most transcriptome studies used few replicates and can only identify the genes with the largest changes in expression, thus lacking the ability to detect differences at the level of biological significance.[4] Some studies have also shown that different methods for detecting genes with differential expression lack sufficient statistical power and have a certain false positive rate and false negative rate.[5] Therefore, increasing the sample size and seeking new analysis strategies are crucial for overcoming the limitations of traditional transcriptome analysis.

Machine learning (ML), a new big data analysis method, can effectively fit complex data and accurately identify samples and genes. Due to the high flexibility of ML algorithms, it is possible to use them in complex omics data analysis.[6] At present, many ML algorithms are being widely applied in this field. ML algorithms are used in biological modeling, and their performance is better than that of traditional mathematical models.[7] Moreover, ML has proven effective in cancer prognosis analysis, therapeutic target prediction, and drug target prediction, in which classification functions can be used to discover new biomarkers and new drug targets.[8] In addition, Belgian researchers studied hundreds of children's blood samples and characteristics of the immune system and found that the ML algorithm obtained an arthritis diagnosis accuracy as high as 90%.[9] In animal husbandry, ML has also begun to be used for genomic selection, with significantly better accuracy than traditional methods.[10] ML has also been gradually applied in the study of the economic traits of pigs. There are studies that have used ML to predict daily gain[11] and total number born[12] of pigs, which showed high accuracy. However, due to

Received: May 16, 2022
Revised: July 27, 2022
Accepted: July 29, 2022
Published: August 11, 2022

https://doi.org/10.1021/acs.jafc.2c03339
J. Agric. Food Chem. 2022, 70, 10359−10370

Table 1. Sample Information from Different Sources

source	breed	tissue	fat content index	high group	low group
ours	Landrace, Songliao black pig	adipose, muscle, liver	backfat thickness	18	18
GSE61271	Duroc × Göttingen Minipigs	adipose	obesity index	12	12
GSE144780	Italian Large White	muscle	intramscular fat content	6	6
GSE116951	Iberian pig	muscle	intramscular fat content	6	6
GSE122349	Pulawska breed	adipose	backfat thickness	8	8

the relatively high cost and complex processing requirements of RNA sequencing, ML also faces the problem of a small sample size. The collection of multiple samples with similar experiments can not only preserve biological information but also improve the effectiveness and practicality of gene expression analysis.[5]

Therefore, in this study, transcriptome sequencing data were collected from the major organs of fat deposition in pigs from multiple sources, and the strategy of imputing missing values was applied to unify data from different sources. In addition, eight ML methods were also compared to evaluate the prediction accuracy of ML models, and genes affecting fat deposition were predicted by the best ML method. Meanwhile, the efficiency of traditional differential expression analysis was also compared with that of ML methods.

MATERIALS AND METHODS

Pig Samples and RNA-Seq. RNA-seq data from five sources were used in this study (Table 1). The experimental population used in this study was from a pig breeding farm in Tianjin, China. A total of 500 Landrace (n = 341) and Songliao black (a Chinese breed, n = 159) sows were selected. The backfat thickness (5 cm between the third and fourth ribs) of live pigs (~100 kg body weight) was measured by B-ultrasound in vivo as an index of fat deposition, because the backfat thickness was highly positively correlated with the fat deposition content.[13] For each breed, six individuals with the highest and lowest backfat thicknesses were sampled. Adipose tissue, muscle tissue, and liver tissue samples were collected from these 24 individuals, and 36 samples were selected according to sample quality, including 16 Songliao black pig samples and 20 Landrace pig samples (Table S1). The numbers of samples in the high- and low-fat content groups of each breed were equal.

All animal studies were evaluated and authorized by the Institutional Animal Care and Use Committee (IACUC). The whole procedure for samples collected was carried out in strict accordance with the protocol approved by the IACUC at the China Agricultural University. The IACUC of the China Agricultural University specifically approved this study (permit number DK996).

We extracted RNA from 36 samples and sequenced mRNA using the Illumina HiSeq 2000 sequencing platform after sample processing. IlluQC.pl (NGS QC Toolkit)[14] was used for quality control of the sequenced reads, and HISAT2[15] was used for fast and accurate sequence alignment. Finally, SAMtools[16] and FeatureCounts[17] were used to transform the transcriptome gene expression count file in order to obtain the gene expression profile in each tissue sample.

In addition, we also collected similar transcriptome data from 64 samples in the GEO database from four different sources (Table 1), including adipose tissue samples of 24 Duroc × Gottingen minipigs[18] and 16 Pulawska pigs[19] and muscle tissue samples of 12 Italian Large White pigs[20] and 12 Iberian pigs.[21] The samples were screened and grouped according to their phenotypic information, including the obesity index, intramuscular fat content, and backfat thickness. Samples from each source were divided into two groups (high- and low-fat contents or obesity indexes), and the numbers of samples in the groups were equal.

Data Standardization and Imputation. Data standardization was first carried out to make the five different sources comparable. Each data set was transformed into fragments per kilobase per million mapped fragments (FPKM) values in a unified manner. Then, the data were combined, and the gene names were transformed according to the pig reference genome (Sscrofa11.1). The genes with gene symbols were retained, and the genes with missing rates greater than 20% were excluded. A variety of strategies were implemented to impute the remaining missing values.

Ten imputation methods (MINIMUM,[22] stochastic minimal value (MINPROB),[23] row median (ROWMEDIAN),[24] singular value decomposition (SVD),[25] maximum likelihood estimation (MLE),[26] sequential imputation (IMSEQ),[27] robust sequential imputation (IMPSEQROB),[28] K-nearest neighbor (KNN),[25] sequential KNN (SEQ-KNN),[29] and quantile regression (QR)[23]) were compared. MINIMUM, MINPROB, and ROWMEDIAN are simple and fast because they are the minimum, random, and median values, respectively, to directly replace missing values. SVD, MLE, IMPSEQ, and IMPSEQROB consider the global structure of the gene matrix, decompose the data matrix or minimize the determinant of the covariance and then iteratively impute the missing values. KNN, SEQ-KNN, and QR consider only values near the missing value and impute missing data on the basis of local similarity of the gene expression profile. Four evaluation criteria, the average correlation coefficient between the original value and imputed value (ACC_OI), NRMSE, NRMSE-based SOR, and PSS, were used to evaluate the efficiency of data imputation.[30] After data imputation, the batch effect of five different sources was corrected using the R package combat.[31] PCA and cluster analysis were carried out on the data before and after removing the batch effect to show the batch effect.

Differentially Expressed Gene Analysis. Differential expression analysis was performed on the data after adjusting for batch effects. Limma,[32] which can fit linear models of gene modes to gene expression data in order to detect differential expression, was used to identify the DEGs. In the merged data of five different sources, the groups were the same as previously set. In addition, DEG analysis was also conducted for individual sources. According to the results of differential expression analysis, genes were ranked according to P value, and genes with a P value less than 0.05 were regarded as differentially expressed.

Machine Learning. To further screen for candidate genes affecting fat deposition, we performed ML analysis based on the results of differential expression analysis. The whole process of ML analysis is illustrated in Figure 1. The 500, 1000, 2000, and 3000 genes with the most significant P values from differential expression analysis were chosen as selection features to facilitate better ML model training. Meanwhile, all the genes (6658) detected in the differential expression analysis were also chosen as selection features. In addition, organization type was also added to the data set in the form of a numerical value as the feature vector of the samples. For these five cases, we conducted ML model training and evaluation. To achieve better and faster convergence of the models, normalization and standardization of the constructed data sets were applied; that is, the expression level of each gene was scaled to 0−1, and the variances of all genes were equalized. For the 100 samples, 1000 replications of fourfold cross validation (CV) were carried out to evaluate the ML model. For each instance of CV, 75 and 25 samples were used to build the classification model and to evaluate the accuracy of the model, respectively. The prediction accuracy of the ML model was the rate of correct sample classification in the validation population. We fine-tuned the hyperparameters of the ML model manually to improve the accuracy of the model prediction.

https://doi.org/10.1021/acs.jafc.2c03339
J. Agric. Food Chem. 2022, 70, 10359−10370

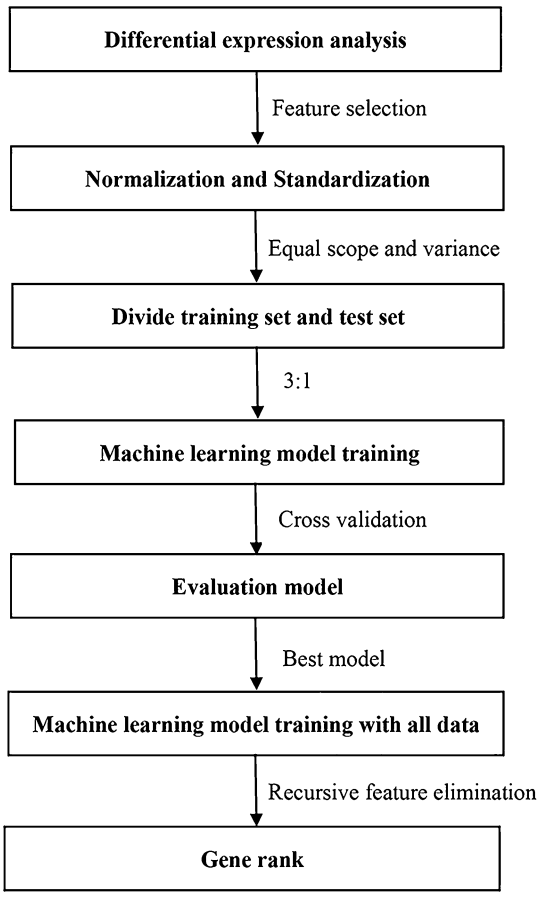

Figure 1. Workflow of machine learning analysis.

To find the ML model that best fits the data in this study, we tested eight commonly used classification models (Linear Support Vector Classification (LinearSVC),[33] Radial Basis Function Kernel Support Vector Machine (RBF SVM),[34] RandomForest,[35] Nearest Neighbors,[36] Gaussian Process,[37] Decision Tree,[38] Neural Network,[39] and AdaBoost[40]). These models are fully supervised ML classification models, including linear, nonlinear, and integrated ensemble methods. According to the accuracy of the ML model, the optimal training features, the optimal training model, and the optimal parameter combination of the model were determined. In addition, for the feature numbers with the highest accuracy, ROC curves were drawn for each of the eight models to further evaluate model quality.

The best model was selected, and all 100 samples were reanalyzed using the model. Each gene was ranked by RFE[41] to screen out the most important genes for model classification. For this study, the higher the rank of the genes based on RFE, the more likely they were to determine whether a sample was classified into a high- or low-fat content group, indicating that these genes play important roles in fat deposition.

The ML model was developed and RFE was applied using the package Scikit-learn V.1.0. All steps were performed using Python V.3.9.6. NumPy V.1.22 and pandas V.1.3.4 were used for data collation and basic statistical calculations, respectively.

Gene Function Analysis. Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses were performed on the top 100 genes screened by the best ML methods, KOBAS[42] was used to conduct enrichment analysis of four features (cellular components, molecular functions, biological processes, and pathways) for the top 100 genes, and a false discovery rate (FDR)-corrected P value less than 0.05 was considered significant. In addition, ClueGO[43] in Cytoscape software[44] was further applied to the top 100 genes to detect relationships between different enrichment pathways. Candidate genes were further selected in combination with the annotation of Sscrofa11.1. In addition, enrichment analysis was also conducted for the top 100 genes screened by P value ranking in DEG analysis and all the DEGs. We also analyzed the expression of candidate genes in 72 different tissue samples (7096 in total) from pigs[45] to discover their unique expression patterns.

RESULTS

Standardization of Gene Expression Data from Different Sources. The numbers of expressed genes in the samples from five sources were 14975, 15455, 9568, 14971, and 14875, among which 8317 genes overlapped (Figure S1). After merging data from the five sources and imputing missing values, 6658 overlapping genes remained, only 5% of all gene expression values were missing, and the distribution of missing values was relatively concentrated (Figure S2). Efficient imputation of missing values can improve the accuracy of subsequent data analysis. Table 2 shows the imputation performance of 10 data imputation methods. Among them, IMSEQROB performed the best; it ranked 1st for all four evaluation criteria, yielding the highest average correlation coefficient (0.99) and the lowest normalized root mean squared error (NRMSE), NRMSE-based sum of ranks (SOR), and Procrustes sum of squared errors (PSS). A similar approach to IMSEQROB, SEQKNN, ranked 2nd in the overall evaluation except in terms of the average correlation coefficient, for which it ranked third. This method yielded a higher NRMSE than the IMPSEQ method, and IMPSEQ was the third best among the 10 data imputation methods. Single-value replacement methods are simple and quick but performed poorly in this study, with MINIMUM and ROWMEDIAN ranking at the bottom of the middle. In all scenarios, MLE performed the worst; it ranked last for each evaluation criterion, generating the lowest average correlation

Table 2. Comparison of Different Methods for Imputing Missing Values

methods	Cor_mean	NRMSE	PSS	SOR	NRMSE_Rank	SOR_Rank	ACC_OI_Rank	PSS_Rank	Rank_Mean
IMPSEQROB	0.9936	0.3640	2.00×10^{-5}	758	1	1	1	1	1
SEQKNN	0.9879	0.5124	4.00×10^{-5}	1290	3	2	2	2	2.25
IMPSEQ	0.9874	0.4118	6.00×10^{-5}	1330	2	3	3	4	3
KNNMETHOD	0.9871	0.5398	5.00×10^{-5}	1490	4	4	4	3	3.75
ROWMEDIAN	0.9554	0.8981	0.00018	2279	5	5	5	5	5
MINIMUM	0.8663	1.0496	0.00154	3148	7	6	6	6	6.25
MINPROB	0.7920	1.0505	0.00215	3489	8	7	8	7	7.5
SVDMETHOD	0.6583	1.0002	0.0034	3608	6	8	9	9	8
QRILC	0.8125	2.9811	0.00283	3668	9	9	7	8	8.25
MLE	0.0124	2351.45	0.01001	4680	10	10	10	10	10

https://doi.org/10.1021/acs.jafc.2c03339
J. Agric. Food Chem. 2022, 70, 10359−10370

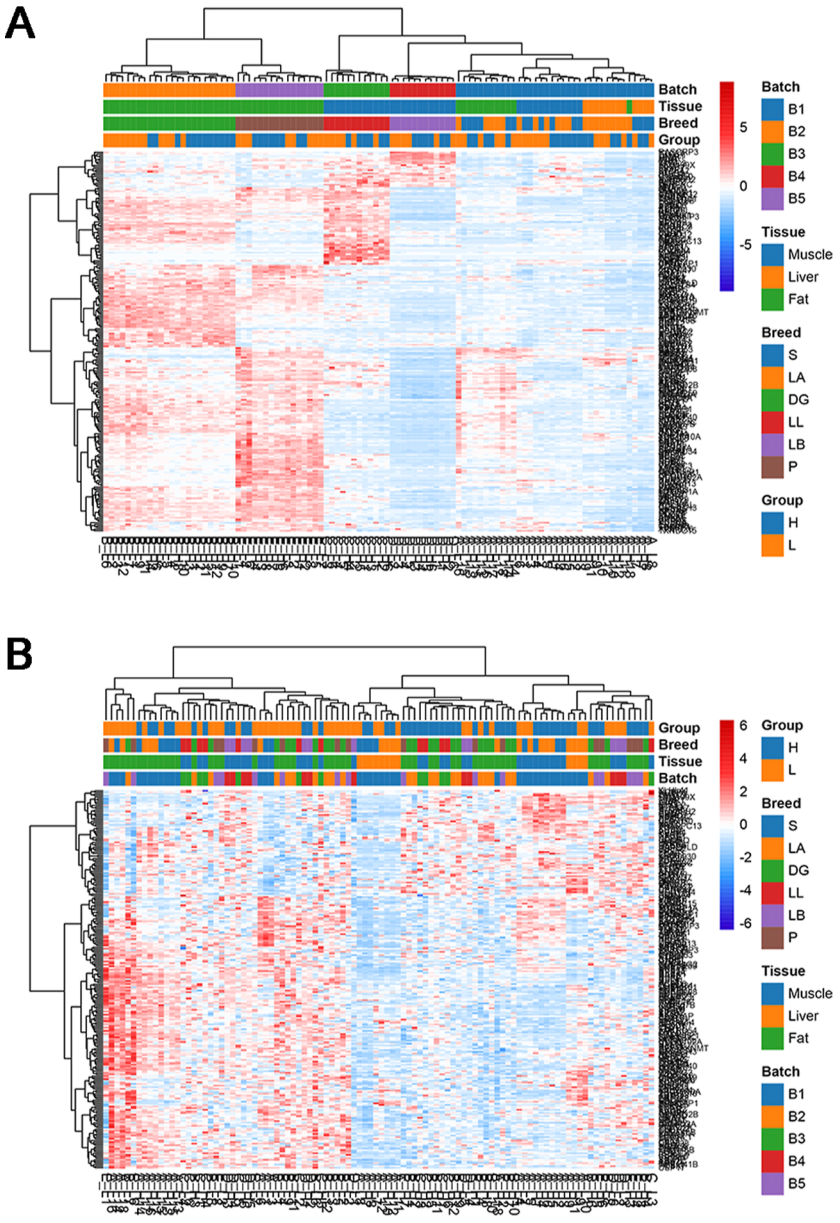

Figure 2. Cluster diagrams of genes before (A) and after (B) removing the batch effect.

coefficient (an extremely low value of 0.01) and highest NRMSE, NRMSE-based SOR, and PSS. Therefore, we chose the IMPSEQROB method for the imputation of missing data.

In the merged data after imputation, a large batch effect was indicated among the five different data sources, as shown in Figure 2, with different data sources clustered onto different branches. Moreover, tissue and breed effects were also detected in the data clustering analysis. Figure 2A further illustrates that all samples were obviously divided into three groups based on principal component analysis, and most of the samples in each group came from the same source, indicating heterogeneity in the data. After correcting for the batch effect by combat, the batch effect, tissue effect, and breed effect were all reduced, as illustrated in Figure 2B. The range of gene expression values in the samples decreased from 10,000 to 8000, PC1 decreased from 57.87 to 45.13% (Figure 3), and the samples were more uniform after standardization.

Analysis of Differentially Expressed Genes in the Merged Data Set. After standardization and removal of the batch effect, 235 differentially expressed genes (DEGs, P value <0.05) were identified by limma. Figure 4 illustrates that the DEGs were mostly downregulated in expression in the high-fat content group and upregulated in expression in the low-fat content group. The 10 genes with the smallest P values are shown in Figure 4B, and the gene with the largest fold change was FASN (logFC = 244.5, P value = 0.016). In addition, 280, 2048, 577, 931, and 2088 DEGs were also identified for individual sources by limma, while no overlap was found among these DEGs (Figure S3), implying that it is difficult to find candidate genes by summarizing results from different data sources. Most of the 235 DEGs in the merged data set overlapped with those from different sources, and only 33 of them could not be found through differential expression analysis of a single source.

Comparison of Eight Machine Learning Models. For different ML models, parameter tuning was carried out, and the optimal parameter combination was selected according to the results of CV (Table S2). Table 3 shows the prediction

https://doi.org/10.1021/acs.jafc.2c03339
J. Agric. Food Chem. 2022, 70, 10359−10370

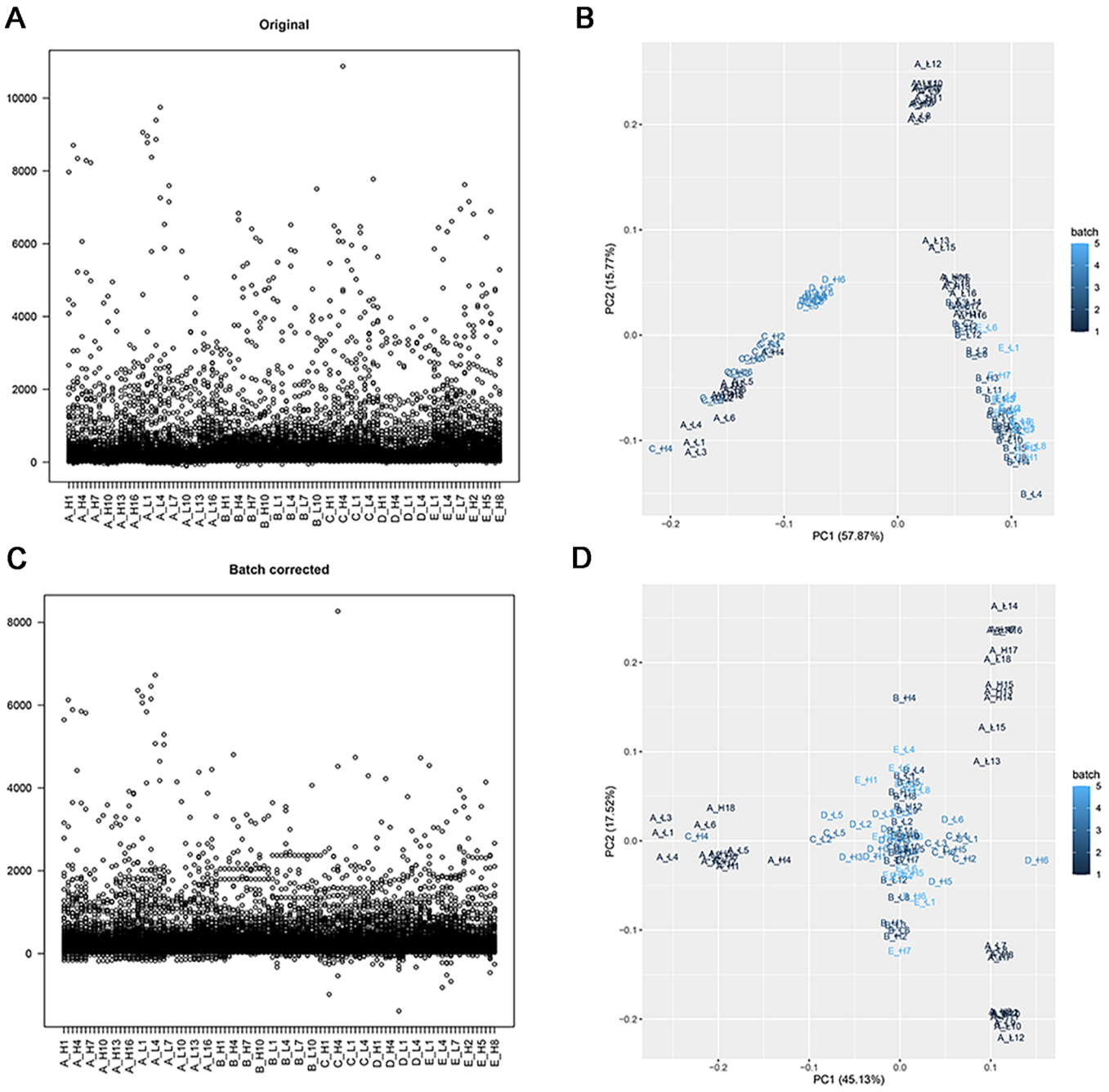

Figure 3. Distribution and PCA before and after removing the batch effect (A,C) is the data distribution and PCA clustering diagram of the original data. (B,D) is the data distribution and PCA cluster graph after removing the batch effect.

accuracies of eight ML models under different characteristic gene scenarios. The accuracies of the eight ML models showed almost the same trend for all tested numbers of selection feature genes. AdaBoost yielded the highest accuracy (more than 90%) in predicting the high- and low-fat content groups, and Nearest Neighbors yielded the lowest accuracy. Similar to that of Nearest Neighbors, the prediction accuracies of Linear SVM, RBF SVM, Gaussian Process, and Neural Net s were lower than 80% in all scenarios, Random Forest yielded accuracies greater than 80%, and Decision Tree performed similarly to AdaBoost. On the other hand, when the top 2000 genes were selected, AdaBoost performed better than when other feature numbers were selected, yielding an average prediction accuracy of 93 and 93.4% in the high- and low-fat content groups, respectively, and the narrowest 95% confidence intervals of 84−100% for both the high- and low-fat content groups. Moreover, the receiver operating characteristic (ROC) curves of the 8 ML methods further illustrated that AdaBoost performed best (Figure 5). The area under the ROC curve (AUC) of AdaBoost was much higher than those of the other models. The ROC curve of the AdaBoost model also had minimal variance, indicating that the model was more stable than the other models for different data sets under CV. Therefore, we selected the AdaBoost model and the top 2000 genes to train all samples and to rank the genes in terms of importance based on recursive feature elimination (RFE).

Identification of Genes with Unique Expression Patterns. We ranked 2000 genes and tissue factors involved

https://doi.org/10.1021/acs.jafc.2c03339
J. Agric. Food Chem. 2022, 70, 10359−10370

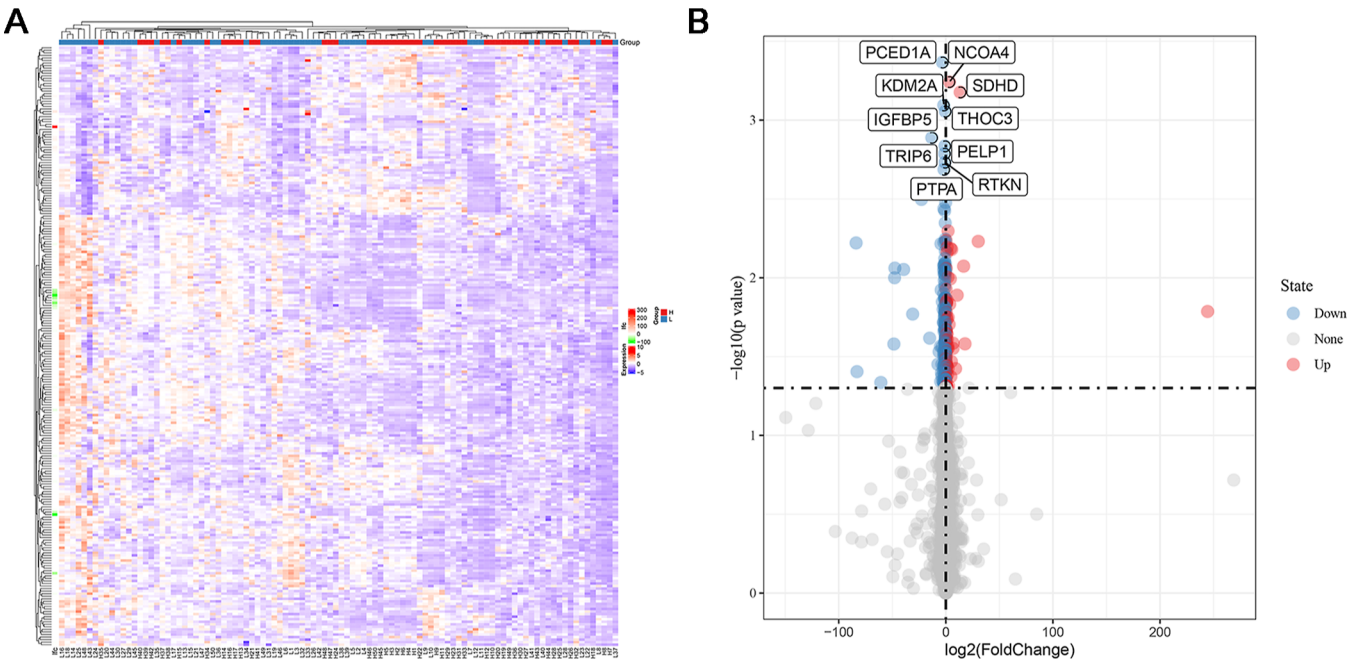

Figure 4. Distribution (A) and Volcanogram (B) of differentially expressed genes (DEGs) (A) Expression distribution of differentially expressed genes and the clustering of samples according to gene expression. (B) Situation of the differential genes. The *x*-axis represents the multiple of difference, which is denoted by log2FoldChange. The larger the absolute value is, the larger the multiple of difference is. The *y*-axis represents the significance of the difference, which is denoted by −log 10 (*P*-value). The larger the value is, the more significant the difference is. The panel shows the names of the top 10 genes with the most significant differences.

in the analysis according to importance by applying RFE. Among the top 100 genes with the highest importance, only 16 were differentially expressed (Figure 6), implying that ML is quite different from traditional differential expression analysis, and the top three genes were EFCAB7, ZDHHC18, and LRPPRC (Table S3). In addition, the functional enrichment of the top 100 genes screened by ML was quite different from that of all DEGs or the top 100 DEGs. There were no common enrichment items, and the DEGs and top 100 differentially expressed enrichment items were not directly related to fat development (Figures 7 and S4).

Through functional enrichment analysis of the top 100 genes screened by ML, the two most significant items directly related to the formation of fat were identified, namely, ether lipid metabolism and lipid catabolic process, which were associated with the genes *PLA2G6*, *PLA2G7*, *PLD4*, and *PLD1* and *PLA2G6, PLCB1, PLD1*, and *PLA2G7*, respectively (Figure 7). Additionally, several gene groups were also involved in lipase activity and energy metabolism. Finally, 12 genes (*IFIT1*, *ZDHHC18*, *FASN*, *PLA2G6*, *PLA2G7*, *PLCB1*, *PLD4*, *PLCG2*, *PLD1*, *APOA1*, *APOD*, and *APOOL*) were found to be associated with fat growth and development and could be candidate genes for the regulation of fat deposition (Table S3). Among them, *ZDHHC18*, *IFIT1*, and *FASN* ranked 2nd, 4th, and 10th, respectively, in the ML model (Table S3). Figure 8 further illustrates the expression of these 12 candidate genes in 12 main tissues from 72 samples of pigs. *FASN* and *APOD* were specifically expressed in adipose tissue. *APOA1* was specifically expressed in the liver. In muscle tissue, the expression levels of these candidate genes were very low.

DISCUSSION

Comprehensive analysis of data is considered a key method for extracting the most effective information from different genomic data sets, which is conducive to the discovery of important biological phenomena.[46] At present, there are two different comprehensive analysis strategies: meta-analysis and data combination. When there is large heterogeneity among original studies and the number of studies is small, random effects cannot be fully considered in the model, resulting in invalid conclusions with type I error and leading to low consistency among studies.[47] This point was confirmed by our findings, in which the DEGs from five different sources showed poor consistency (Figure S3). Therefore, such meta-analysis is not applicable to these data sets. The data combination method involves combining samples from different sources to enlarge the data set and then analyzing the newly combined data set.[48] The advantages of data combination over the meta-analysis method mainly lie in the greater statistical significance of the results obtained by analyzing the combined large sample sets and the more rigorous inference results.[49] Our results confirmed that combining data from five different sources yielded the most DEGs obtained through single-source analysis, and no common DEGs were found among single-source analyses even though many DEGs were identified. In addition, in order to expand the sample size as much as possible to meet the training requirements of machine learning, we put three kinds of important tissue samples directly related to fat deposition in the same data set for joint analysis.

When combining data, some key issues must be solved to unify the data, for example, batch effects and missing values. In this study, batch and tissue type effects were found to influence the uniformity of the data. Many studies have shown that ComBat adjustment of data results in improved statistical power and control of false positives in differential expression analysis compared to those of data adjustment by other available methods.[50] The empirical Bayes method in ComBat was adopted to eliminate the effect of covariates for batch

https://doi.org/10.1021/acs.jafc.2c03339
J. Agric. Food Chem. 2022, 70, 10359−10370

Table 3. Comparison of the Accuracy and Confidence Interval (CI) of Eight Machine Learning Models with Different Feature Numbers on Fat Content Classification in Four-Fold Cross Validation of 1000 Replicates[a]

feature		accuracy and 95% CI		accuracy and 95% CI
Top500 genes	linear SVM	H: 75.3 (56.0−88.0)	Gaussian Process	H: 70.3 (52.0−84.0)
		L: 75.5 (60.0−88.1)		L: 69.9 (52.0−84.0)
	RBF SVM	H: 70.9 (44.0−76.0)	Decision Tree	H: 90.8 (80.0−100.0)
		L: 70.8 (44.0−76.0)		L: 90.8 (80.0−100.0)
	random forest	H: 87.4 (76.0−100.0)	Neural Net	H: 78.1 (64.0−92.0)
		L: 87.7 (76.0−100.0)		L: 77.0 (60.0−92.0)
	nearest neighbors	H: 62.6 (44.0−80.0)	AdaBoost	H: 92.9 (80.0−100.0)
		L: 62.7 (44.0−80.0)		L: 92.9 (84.0−100.0)
Top1000 genes	linear SVM	H: 74.1 (56.0−92.0)	Gaussian Process	H: 68.4 (52.0−84.0)
		L: 74.4 (59.9−88.0)		L: 64.4 (48.0−80.0)
	RBF SVM	H: 66.7 (52.0−80.0)	Decision Tree	H: 92.5 (80.0−100.0)
		L: 66.6 (51.9−84.0)		L: 92.4 (80.0−100.0)
	random forest	H: 85.8 (72.0−96.0)	Neural Net	H: 73.6 (56.0−88.0)
		L: 85.8 (68.0−96.0)		L: 74.2 (56.0−84.0)
	nearest neighbors	H: 59.3 (40.0−76.0)	AdaBoost	H: 92.9 (84.0−100.0)
		L: 59.1 (40.0−76.0)		L: 93.1 (84.0−100.0)
Top2000 genes	linear SVM	H: 70.8 (52.0−88.0)	Gaussian Process	H: 64.0 (48.0−76.0)
		L: 70.6 (52.0−88.0)		L: 58.7 (44.0−72.0)
	RBF SVM	H: 62.8 (48.0−80.0)	Decision Tree	H: 92.0 (80.0−100.0)
		L: 62.7 (44.0−80.0)		L: 91.8 (80.0−100.0)
	random forest	H: 86.0 (72.0−100.0)	Neural Net	H: 69.3 (52.0−88.0)
		L: 86.4 (72.0−100.0)		L: 69.3 (52.0−84.0)
	nearest neighbors	H: 56.2 (40.0−72.0)	AdaBoost	H: 93.0 (84.0−100.0)
		L: 56.0 (40.0−72.0)		L: 93.4 (84.0−100.0)
Top3000 genes	linear SVM	H: 66.8 (48.0−84.0)	Gaussian Process	H: 64.8 (48.0−80.0)
		L: 66.8 (48.0−84.0)		L: 64.5 (48.0−80.0)
	RBF SVM	H: 60.1 (44.0−76.0)	Decision Tree	H: 92.0 (80.0−100.0)
		L: 60.1 (44.0−76.0)		L: 92.1 (80.0−100.0)
	random forest	H: 84.7 (68.0−96.0)	Neural Net	H: 64.6 (48.0−84.0)
		L: 84.6 (68.0−96.0)		L: 64.2 (48.0−84.0)
	nearest neighbors	H: 53.7 (36.0−68.0)	AdaBoost	H: 92.4 (80.0−100.0)
		L: 53.4 (36.0−68.0)		L: 92.7 (84.0−100.0)
all 6658 genes	linear SVM	H: 54.9 (36.0−76.0)	Gaussian Process	H: 48.1 (40.0−56.0)
		L: 54.8 (36.0−72.0)		L: 52.2 (44.0−60.0)
	RBF SVM	H: 52.7 (36.0−68.0)	Decision Tree	H: 91.7 (80.0−100.0)
		L: 52.8 (36.0−68.0)		L: 91.8 (80.0−100.0)
	random forest	H: 86.5 (72.0−100.0)	Neural Net	H: 54.3 (36.0−72.0)
		L: 86.6 (72.0−100.0)		L: 53.3 (36.0−72.0)
	nearest neighbors	H: 49.0 (32.0−64.0)	AdaBoost	H: 92.1 (80.0−98.0)
		L: 48.9 (32.0−64.0)		L: 92.0 (84.0−100.0)

[a]Models are sorted based on the average accuracy of all the groups. H: high-fat group, L: low-fat group.

effect correction. Our results indicated that the batch effect was significantly corrected (Figures 2 and 3). Although a good trial design can reduce batch effects, it is difficult to eliminate systematic bias completely;[51] therefore, we incorporated the tissue type of the sample as a feature into the ML training model to reduce bias. Regarding missing values, we compared the imputation effects of various methods and showed that Global Structure Approach IMPSEQROB ranked first in all evaluations because it fully considers the relationships between genes, which is more in line with biological characteristics. Studies have shown that when processing biological data such as protein expression data, IMPSEQROB has a higher completion effect on missing values, and the distribution of filled data is more similar to that of real data.[30]

There is no perfect ML algorithm that can solve all problems; instead, ML algorithms must be tailored to different data.[52] In this study, we evaluated almost all popular ML algorithms through CV. Compared with SVM and neural network algorithms widely used in the biomedical field, the AdaBoost algorithm performed best in this study, which further indicates that different algorithms should be used for specific problems. The neural network approach has better model complexity and can fit complex data more accurately than other approaches. However, for this study, the sample information complexity was high, but the number of training samples was relatively small, and the samples were not as heterogeneous as tumor samples, so overfitting was easily caused by using an overly complex model. By training different weak classifiers, the AdaBoost algorithm integrates these weak classifiers to form a strong classifier, which can solve the problem of complex structures with high classification accuracy. The subclassifier is a CART decision tree. As a binary classifier, it has a simple structure. It not only has a very fast training speed but also can adapt well to the characteristics of small training sets such as that used in this study. The most important feature is that the whole model is not easy to overfit, and its iteration process has the effect of increasing the margin.[53] The results also indicate that higher model complexity in ML is not necessarily better. For data with a small sample size and complex information, an integrated ML method similar to AdaBoost can be adopted. In addition, we found that the accuracy of the decision tree was also quite high

https://doi.org/10.1021/acs.jafc.2c03339
J. Agric. Food Chem. 2022, 70, 10359−10370

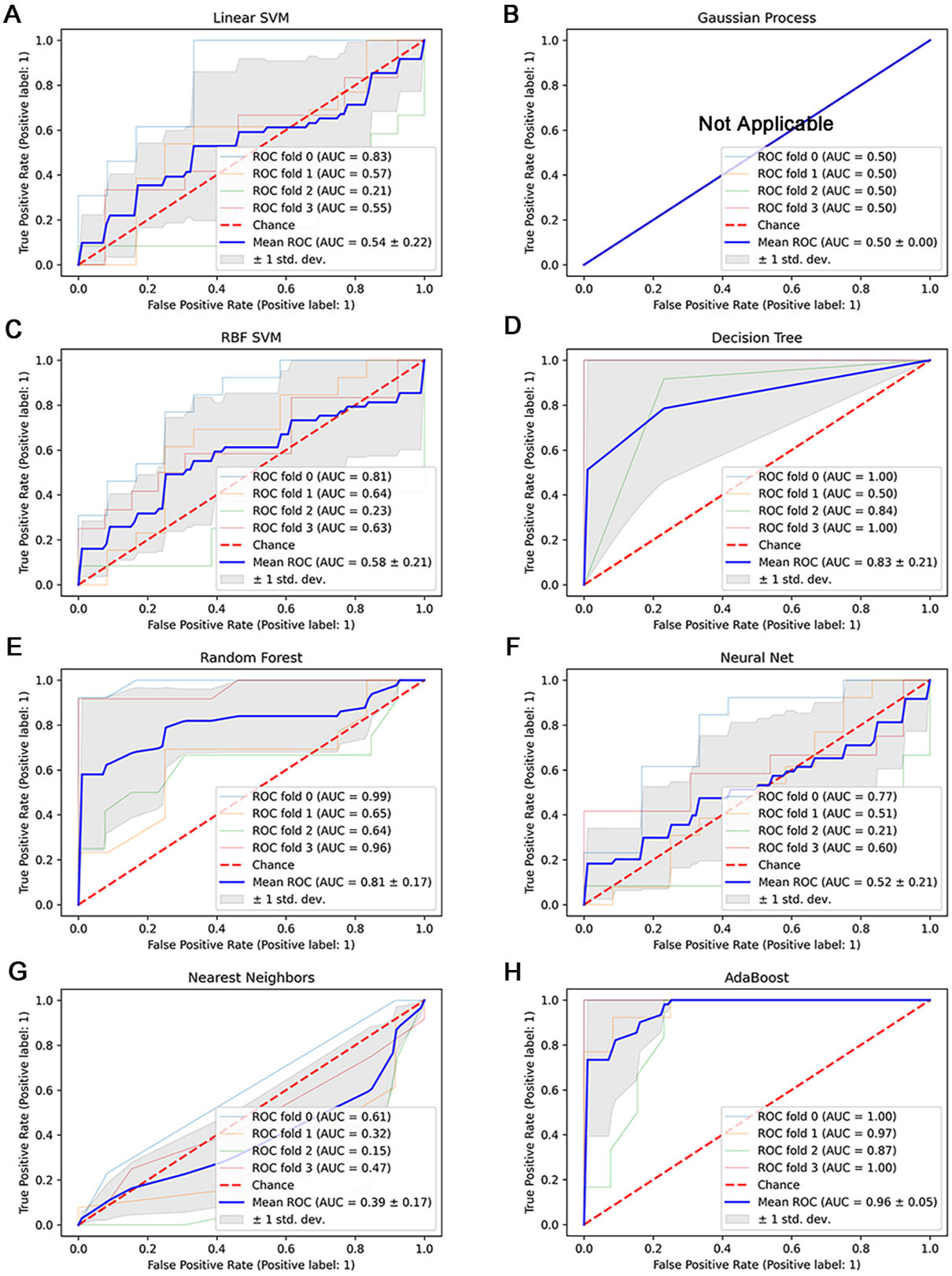

Figure 5. Receiver operating characteristic (ROC) curves of eight machine learning models with 2000 selection feature genes. The figure shows the ROC response of different data sets, created from four-fold cross validation. The blue curve shows the mean value under different conditions, which can represent the average performance of the model.

(>90%), as its algorithm is similar to AdaBoost's basic classifier, but both the accuracy and precision of the decision tree were lower than those of AdaBoost. Therefore, the integration of multiple identical models can not only effectively improve accuracy but also improve precision. In addition, the parameters of the model can directly affect its prediction

https://doi.org/10.1021/acs.jafc.2c03339
J. Agric. Food Chem. 2022, 70, 10359−10370

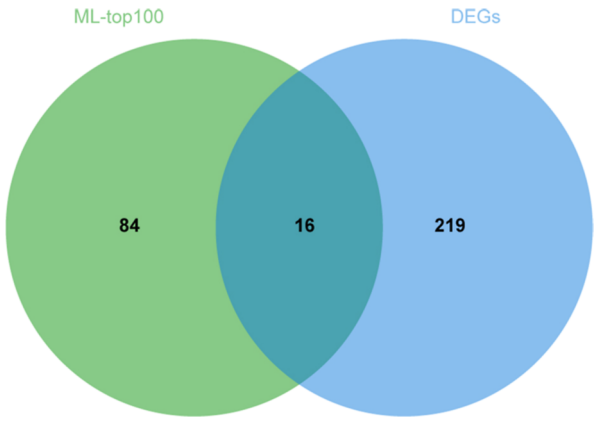

Figure 6. Venn map of differential expression analysis and machine learning The figure shows the distribution of the top 100 genes by machine learning and differentially expressed genes.

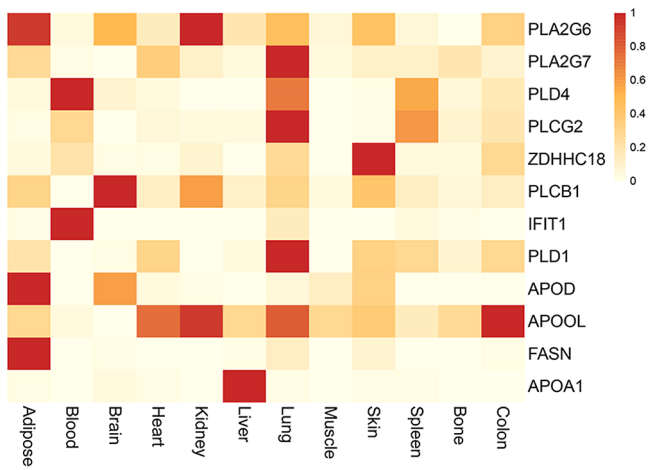

Figure 8. Expression of candidate genes in different tissues of pigs. The expression level of genes is the standardized TPM value, and the expression level in different tissues is the mean value of different samples, then scaled to 0−1.

accuracy.[54] In this study, only a few parameters were manually adjusted for each model, and almost all models performed best under their default parameters (Table S4). Other parameters were not tested one by one in this study, and the selected parameters may not be the best ones. However, in this study, AdaBoost's AUC was 0.96, almost as high as expected (Table 3 and Figure 5).

Sample feature selection is needed for training ML models, but there is currently no unified screening standard for sample features. In this study, the top 500, 1000, 2000, and 3000 genes based on differential expression analysis were used; in addition, all genes were also used for feature selection. Our results showed that the highest accuracy was obtained for ML with the top 2000 genes (Table 3). We noticed that if feature selection is not carried out, almost all ML models have the lowest prediction accuracy because the feature number is far larger than the sample number, which indeed causes serious overfitting. Therefore, implementing differential expression analysis for feature selection is straightforward and useful. On the other hand, the determination of candidate genes in ML is different from that in traditional differential expression analysis. The genes were ranked in terms of importance by using RFE and repeated model building, and the most important genes contributed more to ML model classification. Differential expression analysis can identify candidate genes only by significance or multiple differences, with a certain rate of false positives. RFE is not only suitable for multiple models but also can accurately rank each gene, providing a new method for screening important genes, and it has been proven to be effective in relevant studies.[55]

The enrichment items of the top 100 genes screened by the AdaBoost algorithm showed high relevance to fat deposition, indicating that this method is effective in screening candidate genes. We further screened 12 candidate genes, all of which have been shown to be involved in regulating fat deposition. *PLA2G6* and *PLA2G7* catalyze the hydrolysis of phospholipids (PLs) to generate fatty acids, and their abnormal expression

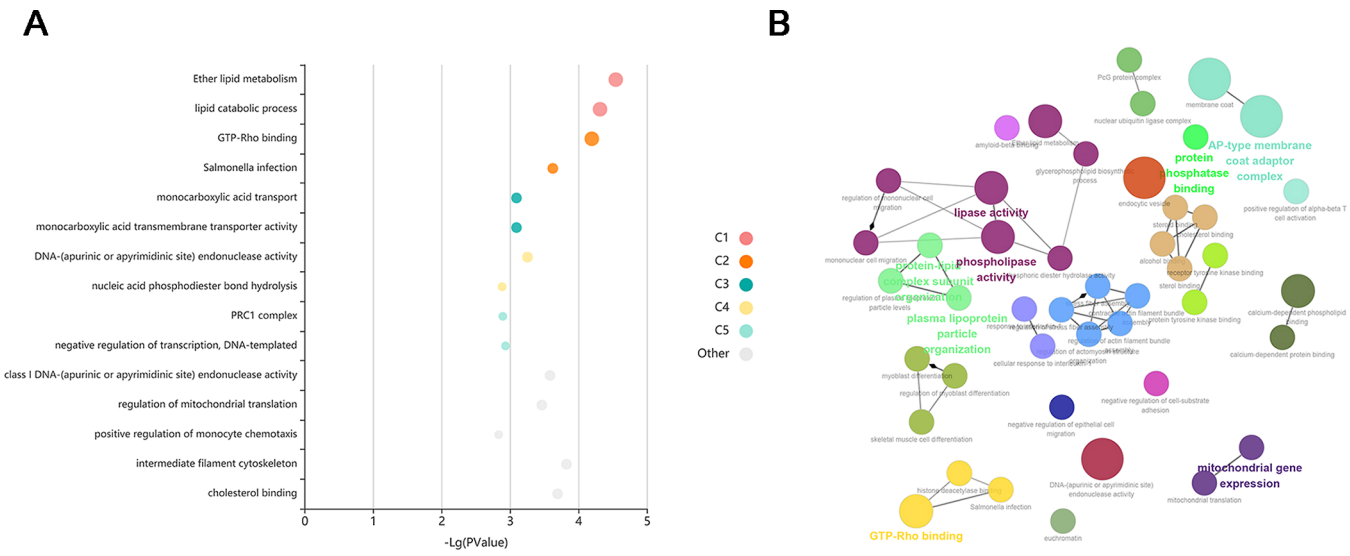

Figure 7. Gene enrichment analysis of top 100 genes in machine learning (A). The enrichment of top 100 genes in machine learning. The *y*-axis represents the gene enriched entries according to the GO and KEGG analyses, and the *x*-axis represents −Lg(*P* value) of the enriched entries or path. The color of the bar is the same as the color in the circular network. (B) Gene-enriched items and the relationships between the items according to ClueGO. The lines between the dots indicate the presence of common genes between items.

https://doi.org/10.1021/acs.jafc.2c03339
J. Agric. Food Chem. 2022, 70, 10359−10370

can cause dysregulated lipid metabolism.[56] *PLD1* regulates cytosolic lipid droplet formation, and increased expression of *PLD1* increases lipid droplet formation.[57] Phospholipase D4 (*PLD4*) affects phospholipase activity in mice, which in turn affects fat deposition.[58] Deficiency of PLCB1 leads to myotonic dystrophy because pi-PLC β1 is involved in adipogenesis through a double phase mechanism.[59] Studies have shown that *PLD1*, *PLCB1*, and *ZDHHC18* are highly relevant to lipid phenotypes.[60] A decrease in lipid droplet content was accompanied by a decrease in *IFIT1* expression, and *IFIT1* affects the metabolism of fatty acids by regulating fat oxidation.[61] Fatty acid synthase (*FASN*) is a key enzyme in the synthesis of fatty acids in mammals and predominantly generates straight-chain fatty acids using acetyl-CoA as the initiating substrate. It is directly involved in the regulation of fat formation.[62] Apolipoprotein A1 (*ApoA1*) has been verified to play a vital role in modulating lipid metabolism and homeostasis both in plasma and in cells, consequently affecting fat deposition.[63] Similar to *APOA1*, an important aspect of *APOD*'s role in lipid metabolism appears to involve the transport of arachidonic acid and the modulation of eicosanoid production and delivery in metabolic tissues.[64] Overexpression of *APOOL* led to fragmentation of mitochondria, a reduced basal oxygen consumption rate, and altered crista morphology. Its expression is closely related to energy metabolism.[65]

The overlap between the top 100 genes in the ML analysis and DEGs was less than 20% (Figure 6). The Spearman correlation between gene ranks based on ML and DEG analysis was only 0.046. Therefore, ML is different from differential gene expression analysis. According to the results of functional enrichment analysis of the top 100 genes and DEGs, the two most significant items of the former were directly related to fat deposition, while none of the items of the latter were significantly related to fat development, indicating that ML can yield more convincing findings (Figures 7 and S4). In addition, ML can find useful information that differential expression gene analysis cannot find. Of the 12 candidate genes involved in fat deposition identified by ML, only three were DEGs, implying a high false positive rate of DEG analysis, as pointed out in many other studies. In contrast to the single-gene scope of DEG analysis, ML can consider a large number of genes simultaneously and analyze them as a whole. The nine candidate genes (*PLA2G6*, *PLA2G7*, *PLCB1*, *PLD4*, *PLCG2*, *PLD1*, *APOA1*, *APOD*, and *APOOL*) belong to the same enrichment item or pathway, and most belong to the same family of protein-coding genes. However, they were adjacent in ML rankings (Table S4), suggesting that ML was able to classify them as genes with similar effects. The use of AdaBoost ML to group genes has shown extraordinary biometric capabilities that traditional statistical analyses such as differential expression analysis do not. This is also a unique advantage of ML, but the biological algorithm principle needs to be further explored. Among the 12 candidate genes found by ML, *FASN*, *APOD*, and *APOA1* were highly expressed only in adipose tissue and liver tissue (Figure 8), showing strong tissue specificity, which further indicated that they were closely related to the occurrence of fat deposition and played a more important role than other genes; they could be verified in future studies.

In conclusion, machine learning can efficiently analyze large data sets, and can find useful information that differential expression gene analysis cannot find. This research strategy can provide ideas for the merged analysis of large data sets. According to the results of machine learning analysis, 12 genes including FASN, APOD, and APOA1 may be involved in the regulation of fat deposition in pigs, which lays a foundation for further research on the molecular regulation mechanism behind fat deposition in pigs. The results can provide a reference for the genetic improvement of pork quality traits.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jafc.2c03339.

Sample information, adjustment of model parameters, top 100 genes selected by AdaBoost, ranking and enrichment entries of some important genes based on machine learning, expressed genes and differentially expressed genes in the samples from 5 sources, distribution of missing data, gene enrichment analysis of DEGs, and top100 genes in differential expression gene analysis (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Chuduan Wang − *National Engineering Laboratory for Animal Breeding, Laboratory of Animal Genetics, Breeding and Reproduction, Ministry of Agriculture, College of Animal Science and Technology, China Agricultural University, Beijing 100193, China*; Phone: 010-62732731; Email: cdwang@cau.edu.cn

Xiangdong Ding − *National Engineering Laboratory for Animal Breeding, Laboratory of Animal Genetics, Breeding and Reproduction, Ministry of Agriculture, College of Animal Science and Technology, China Agricultural University, Beijing 100193, China*; Phone: 010-62734277; Email: xding@cau.edu.cn

Authors

Huatao Liu − *National Engineering Laboratory for Animal Breeding, Laboratory of Animal Genetics, Breeding and Reproduction, Ministry of Agriculture, College of Animal Science and Technology, China Agricultural University, Beijing 100193, China*; orcid.org/0000-0002-2381-2584

Kai Xing − *National Engineering Laboratory for Animal Breeding, Laboratory of Animal Genetics, Breeding and Reproduction, Ministry of Agriculture, College of Animal Science and Technology, China Agricultural University, Beijing 100193, China*

Yifan Jiang − *National Engineering Laboratory for Animal Breeding, Laboratory of Animal Genetics, Breeding and Reproduction, Ministry of Agriculture, College of Animal Science and Technology, China Agricultural University, Beijing 100193, China*

Yibing Liu − *National Engineering Laboratory for Animal Breeding, Laboratory of Animal Genetics, Breeding and Reproduction, Ministry of Agriculture, College of Animal Science and Technology, China Agricultural University, Beijing 100193, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jafc.2c03339

Funding

This work was supported by grants for the National Key Research and Development Project (2019YFE0106800), the China Agriculture Research System of MOF and MARA-(CARS-35), the National Natural Science Foundation of China (32070568), and the Beijing Innovation Consortium of Agriculture Research System (BAIC02-2020).

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Hugo, A.; Roodt, E. Significance of porcine fat quality in meat technology: A review. *Food Rev. Int.* **2007**, *23*, 175−198.
(2) Kobayashi, T.; Zhang, H. X.; Tang, W. W. C.; Irie, N.; Withey, S.; Klisch, D.; Sybirna, A.; Dietmann, S.; Contreras, D. A.; Webb, R.; et al. Principles of early human development and germ cell program from conserved model systems. *Nature* **2017**, *546*, 416−420.
(3) Gleeson, M. Basic metabolism I: fat. *Surgery* **2005**, *23*, 83.
(4) Li, Y.; Yang, H.; Zhang, H.; Liu, Y.; Shang, H.; Zhao, H.; Zhang, T.; Tu, Q. Decode-seq: a practical approach to improve differential gene expression analysis. *Genome Biol.* **2020**, *21*, 66.
(5) Rajkumar, A. P.; Qvist, P.; Lazarus, R.; Lescai, F.; Ju, J.; Nyegaard, M.; Mors, O.; Børglum, A. D.; Li, Q.; Christensen, J. H. Experimental validation of methods for differential gene expression analysis and sample pooling in RNA-seq. *BMC Genom.* **2015**, *16*, 548.
(6) Reel, P. S.; Reel, S.; Pearson, E.; Trucco, E.; Jefferson, E. Using machine learning approaches for multi-omics data analysis: A review. *Biotechnol. Adv.* **2021**, *49*, 107739.
(7) Booth, T. C.; Williams, M.; Luis, A.; Cardoso, J.; Ashkan, K.; Shuaib, H. Machine learning and glioma imaging biomarkers. *Clin. Radiol.* **2020**, *75*, 20−32.
(8) Huang, S. J.; Cai, N. G.; Pacheco, P. P.; Narandes, S.; Wang, Y.; Xu, W. N. Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. *Cancer Genomics Proteomics* **2018**, *15*, 41−51.
(9) Van Nieuwenhove, E.; Lagou, V.; Van Eyck, L.; Dooley, J.; Bodenhofer, U.; Roca, C.; Vandebergh, M.; Goris, A.; Humblet-Baron, S.; Wouters, C.; et al. Machine learning identifies an immunological pattern associated with multiple juvenile idiopathic arthritis subtypes. *Ann. Rheum. Dis.* **2019**, *78*, 617−628.
(10) Waldmann, P.; Pfeiffer, C.; Mészáros, G. Sparse Convolutional Neural Networks for Genome-Wide Prediction. *Front. Genet.* **2020**, *11*, 25.
(11) Tusell, L.; Bergsma, R.; Gilbert, H.; Gianola, D.; Piles, M. Machine Learning Prediction of Crossbred Pig Feed Efficiency and Growth Rate From Single Nucleotide Polymorphisms. *Front. Genet.* **2020**, *11*, 567818.
(12) Wang, X.; Shi, S. L.; Wang, G. J.; Luo, W. X.; Wei, X.; Qiu, A.; Luo, F.; Ding, X. D. Using machine learning to improve the accuracy of genomic prediction of reproduction traits in pigs. *J. Anim. Sci. Biotechnol.* **2022**, *13*, 60.
(13) Suzuki, K.; Inomata, K.; Katoh, K.; Kadowaki, H.; Shibata, T. Genetic correlations among carcass cross-sectional fat area ratios, production traits, intramuscular fat, and serum leptin concentration in Duroc pigs. *J. Anim. Sci.* **2009**, *87*, 2209−2215.
(14) Patel, R. K.; Jain, M. NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data. *PLoS One* **2012**, *7*, No. e30619.
(15) Kim, D.; Langmead, B.; Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **2015**, *12*, 357−360.
(16) Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; Proc, G. P. D. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078−2079.
(17) Liao, Y.; Smyth, G. K.; Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **2014**, *30*, 923−930.
(18) Kogelman, L. J. A.; Cirera, S.; Zhernakova, D. V.; Fredholm, M.; Franke, L.; Kadarmideen, H. N. Identification of co-expression gene networks, regulatory genes and pathways for obesity based on adipose tissue RNA Sequencing in a porcine model. *BMC Med. Genomics* **2014**, *7*, 57.
(19) Piórkowska, K.; Małopolska, M.; Ropka-Molik, K.; Szyndler-Nędza, M.; Wiechniak, A.; Żukowski, K.; Lambert, B.; Tyra, M. Evaluation of SCD, ACACA and FASN Mutations: Effects on Pork Quality and Other Production Traits in Pigs Selected Based on RNA-Seq Results. *Animals* **2020**, *10*, 123.
(20) Zappaterra, M.; Gioiosa, S.; Chillemi, G.; Zambonelli, P.; Davoli, R. Dissecting the Gene Expression Networks Associated with Variations in the Major Components of the Fatty Acid Semimembranosus Muscle Profile in Large White Heavy Pigs. *Animals* **2021**, *11*, 628.
(21) Muñoz, M.; García-Casco, J. M.; Caraballo, C.; Fernández-Barroso, M. A.; Sánchez-Esquiliche, F.; Gómez, F.; Rodríguez, M. D. C.; Silió, L. Identification of Candidate Genes and Regulatory Factors Underlying Intramuscular Fat Content Through Longissimus Dorsi Transcriptome Analyses in Heavy Iberian Pigs. *Front. Genet.* **2018**, *9*, 608.
(22) Jiang, Y.; Sun, A.; Sun, Y.; Zhao, W.; Ying, H.; Sun, X.; Yang, B.; Xing, W.; Sun, L.; Ren, B.; et al. Proteomics identifies new therapeutic targets of early-stage hepatocellular carcinoma. *Nature* **2019**, *567*, 257−261.
(23) Lazar, C. *ImputeLCMD: A Collection of Methods for Left-Censored Missing Data Imputation*, 2015.
(24) Meyer, D.; Dimitriadou, E. et al. Misc Functions of the Department of Statistics. *ProbabilityTheory Group (Formerly: E1071)*; TU Wien, 2015.
(25) Troyanskaya, O.; Cantor, M.; Sherlock, G.; Brown, P.; Hastie, T.; Tibshirani, R.; Botstein, D.; Altman, R. B. Missing value estimation methods for DNA microarrays. *Bioinformatics* **2001**, *17*, 520−525.
(26) Xiao, J.; Xu, Q.; Wu, C.; Gao, Y.; Hua, T.; Xu, C. Performance Evaluation of Missing-Value Imputation Clustering Based on a Multivariate Gaussian Mixture Model. *PLoS One* **2016**, *11*, No. e0161112.
(27) Verboven, S.; Branden, K. V.; Goos, P. Sequential imputation for missing values. *Comput. Biol. Chem.* **2007**, *31*, 320−327.
(28) Branden, K. V.; Verboven, S. Robust data imputation. *Comput. Biol. Chem.* **2009**, *33*, 7−13.
(29) Kim, K. Y.; Kim, B. J.; Yi, G. S. Reuse of imputed data in microarray analysis increases imputation efficiency. *BMC Bioinf.* **2004**, *5*, 160.
(30) Wang, S.; Li, W.; Hu, L.; Cheng, J.; Yang, H.; Liu, Y. NAguideR: performing and prioritizing missing value imputations for consistent bottom-up proteomic analyses. *Nucleic Acids Res.* **2020**, *48*, No. e83.
(31) (a) Leek, J. T.; Scharpf, R. B.; Bravo, H. C.; Simcha, D.; Langmead, B.; Johnson, W. E.; Geman, D.; Baggerly, K.; Irizarry, R. A. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* **2010**, *11*, 733−739. (b) Johnson, W. E.; Li, C.; Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **2007**, *8*, 118−127. (c) Müller, C.; Schillert, A.; Röthemeier, C.; Trégouët, D. A.; Proust, C.; Binder, H.; Pfeiffer, N.; Beutel, M.; Lackner, K. J.; Schnabel, R. B.; et al. Removing Batch Effects from Longitudinal Gene Expression - Quantile Normalization Plus ComBat as Best Approach for Microarray Transcriptome Data. *PLoS One* **2016**, *11*, No. e0156594.
(32) Ritchie, M. E.; Phipson, B.; Wu, D.; Hu, Y. F.; Law, C. W.; Shi, W.; Smyth, G. K. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **2015**, *43*, No. e47.
(33) Fan, R. E.; Hsieh, C.-J.; et al. Liblinear: a library for large linear classification. *J. Mach. Learn. Res.* **2008**, *9*, 1871−1874.

https://doi.org/10.1021/acs.jafc.2c03339
J. Agric. Food Chem. 2022, 70, 10359−10370

(34) Pirooznia, M.; Deng, Y. SVM Classifier - a comprehensive java interface for support vector machine classification of microarray data. *BMC Bioinf.* **2006**, *7*, S25.
(35) Statnikov, A.; Wang, L.; Aliferis, C. F. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinf.* **2008**, *9*, 319.
(36) Goldberger, J.; Roweis, S.; Hinton, G. E.; Salakhutdinov, R. R. Neighbourhood Components Analysis. *Adv. Neural Inf. Process. Syst.* **2005**, *17*, 513−520.
(37) Bajer, L.; Pitra, Z.; Repický, J.; Holeňa, M. Gaussian Process Surrogate Models for the CMA Evolution Strategy. *Evol. Comput.* **2019**, *27*, 665−697.
(38) Che, D.; Liu, Q.; Rasheed, K.; Tao, X. Decision tree and ensemble learning algorithms with their applications in bioinformatics. *Adv. Exp. Med. Biol.* **2011**, *696*, 191−199.
(39) Kriegeskorte, N.; Golan, T. Neural network models and deep learning. *Curr. Biol.* **2019**, *29*, R231−R236.
(40) Freund, Y.; Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119−139.
(41) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(42) Bu, D.; Luo, H.; Huo, P.; Wang, Z.; Zhang, S.; He, Z.; Wu, Y.; Zhao, L.; Liu, J.; Guo, J.; et al. KOBAS-i: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis. *Nucleic Acids Res.* **2021**, *49*, W317−W325.
(43) Bindea, G.; Mlecnik, B.; Hackl, H.; Charoentong, P.; Tosolini, M.; Kirilovsky, A.; Fridman, W. H.; Pagès, F.; Trajanoski, Z.; Galon, J. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **2009**, *25*, 1091−1093.
(44) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **2003**, *13*, 2498−2504.
(45) Pan, Z.; Yao, Y.; Yin, H.; Cai, Z.; Wang, Y.; Bai, L.; Kern, C.; Halstead, M.; Chanthavixay, G.; Trakooljul, N.; et al. Pig genome functional annotation enhances the biological interpretation of complex traits and human disease. *Nat. Commun.* **2021**, *12*, 5848.
(46) Rhodes, D. R.; Chinnaiyan, A. M. Integrative analysis of the cancer transcriptome. *Nat. Genet.* **2005**, *37*, S31−S37.
(47) (a) Nakagawa, S.; Noble, D. W.; Senior, A. M.; Lagisz, M. Meta-evaluation of meta-analysis: ten appraisal questions for biologists. *BMC Biol* **2017**, *15*, 18. (b) Guolo, A.; Varin, C. Random-effects meta-analysis: the number of studies matters. *Stat. Methods Med. Res.* **2017**, *26*, 1500−1518.
(48) Lazar, C.; Meganck, S.; Taminau, J.; Steenhoff, D.; Coletta, A.; Molter, C.; Weiss-Solis, D. Y.; Duque, R.; Bersini, H.; Nowe, A. Batch effect removal methods for microarray gene expression data integration: a survey. *Briefings Bioinf.* **2013**, *14*, 469−490.
(49) Hornung, R.; Boulesteix, A. L.; Causeur, D. Combining location-and-scale batch effect adjustment with data cleaning by latent factor adjustment. *BMC Bioinf.* **2016**, *17*, 27.
(50) Zhang, Y.; Parmigiani, G.; Johnson, W. E. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR: Genomics Bioinf.* **2020**, *2*, lqaa078.
(51) (a) Chen, C.; Grennan, K.; Badner, J.; Zhang, D. D.; Gershon, E.; Jin, L.; Liu, C. Y. Removing Batch Effects in Analysis of Expression Microarray Data: An Evaluation of Six Batch Adjustment Methods. *PLoS One* **2011**, *6*, No. e17238. (b) Nyamundanda, G.; Poudel, P.; Patil, Y.; Sadanandam, A. A Novel Statistical Method to Diagnose, Quantify and Correct Batch Effects in Genomic Studies. Sci Rep-Uk 2017, 7.10849 DOI: 10.1038/s41598-017-11110-6
(52) Mirza, B.; Wang, W.; Wang, J.; Choi, H.; Chung, N. C.; Ping, P. Machine Learning and Integrative Analysis of Biomedical Big Data. *Genes* **2019**, *10*, 87.
(53) Gao, W.; Zhou, Z. H. On the doubt about margin explanation of boosting. *Artif. Intell.* **2013**, *203*, 1−18.
(54) Nematzadeh, S.; Kiani, F.; Torkamanian-Afshar, M.; Aydin, N. Tuning hyperparameters of machine learning algorithms and deep neural networks using metaheuristics: A bioinformatics study on biomedical and biological cases. *Comput. Biol. Chem.* **2022**, *97*, 107619.
(55) (a) Wang, T. X.; Shao, W.; Huang, Z.; Tang, H. X.; Zhang, J.; Ding, Z. M.; Huang, K. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nat. Commun.* **2021**, *12*, 3445. (b) Pinal-Fernandez, I.; Casal-Dominguez, M.; Derfoul, A.; Pak, K.; Miller, F. W.; Milisenda, J. C.; Grau-Junyent, J. M.; Selva-O'Callaghan, A.; Carrion-Ribas, C.; Paik, J. J.; et al. Machine learning algorithms reveal unique gene expression profiles in muscle biopsies from patients with different types of myositis. *Ann. Rheum. Dis.* **2020**, *79*, 1234−1242.
(56) (a) Alecu, I.; Bennett, S. A. L. Dysregulated Lipid Metabolism and Its Role in alpha-Synucleinopathy in Parkinson's Disease. *Front. Neurosci.* **2019**, *13*, 328. (b) Pan, G. Z.; Kou, L. J.; Wu, Y.; Hu, Y. J.; Lin, X. S.; Guo, J. R.; Ren, X. X.; Zhang, Y. Regulation of lipoprotein-associated phospholipase A2 silencing on myocardial fibrosis in mice with coronary atherosclerosis. *Biochem. Biophys. Res. Commun.* **2019**, *514*, 450−455.
(57) Andersson, L.; Boström, P.; Ericson, J.; Rutberg, M.; Magnusson, B.; Marchesan, D.; Ruiz, M.; Asp, L.; Huang, P.; Frohman, M. A.; et al. PLD1 and ERK2 regulate cytosolic lipid droplet formation. *J. Cell Sci.* **2006**, *119*, 2246−2257.
(58) Otani, Y.; Yamaguchi, Y.; Sato, Y.; Furuichi, T.; Ikenaka, K.; Kitani, H.; Baba, H. PLD4 Is Involved in Phagocytosis of Microglia: Expression and Localization Changes of PLD4 Are Correlated with Activation State of Microglia. *PLoS One* **2011**, *6*, No. e27544.
(59) Ratti, S.; Mongiorgi, S.; Ramazzotti, G.; Follo, M. Y.; Mariani, G. A.; Suh, P. G.; McCubrey, J. A.; Cocco, L.; Manzoli, L. Nuclear Inositide Signaling Via Phospholipase C. *J. Cell. Biochem.* **2017**, *118*, 1969−1978.
(60) Kathiresan, S.; Manning, A. K.; Demissie, S.; D'Agostino, R. B.; Surti, A.; Guiducci, C.; Gianniny, L.; Burtt, N. P.; Melander, O.; Orho-Melander, M.; et al. A genome-wide association study for blood lipid phenotypes in the Framingham Heart Study. *BMC Med. Genet.* **2007**, *8*, S17.
(61) Monson, E. A.; Crosse, K. M.; Das, M.; Helbig, K. J. Lipid droplet density alters the early innate immune response to viral infection. *PLoS One* **2018**, *13*, No. e0190597.
(62) Wallace, M.; Green, C. R.; Roberts, L. S.; Lee, Y. M.; McCarville, J. L.; Sanchez-Gurmaches, J.; Meurs, N.; Gengatharan, J. M.; Hover, J. D.; Phillips, S. A.; et al. Enzyme promiscuity drives branched-chain fatty acid synthesis in adipose tissues. *Nat. Chem. Biol.* **2018**, *14*, 1021.
(63) Su, X.; Peng, D. Q. The exchangeable apolipoproteins in lipid metabolism and obesity. *Clin. Chim. Acta* **2020**, *503*, 128−135.
(64) Rassart, E.; Desmarais, F.; Najyb, O.; Bergeron, K. F.; Mounier, C. *Apolipoprotein D Gene*; Elsevier, 2020; Vol. *756*.
(65) Weber, T. A.; Koob, S.; Heide, H.; Wittig, I.; Head, B.; van der Bliek, A.; Brandt, U.; Mittelbronn, M.; Reichert, A. S. APOOL Is a Cardiolipin-Binding Constituent of the Mitofilin/MINOS Protein Complex Determining Cristae Morphology in Mammalian Mitochondria. *PLoS One* **2013**, *8*, No. e63683.