J Appl Biotechnol Rep. 2025 December;12(4):1852-1864

Journal of
Applied Biotechnology Reports

doi 10.30491/jabr.2025.490923.1811

Original Article

Combining Machine Learning Algorithms with Meta-Analysis and WGCNA to Identify Biomarker-Responsive Genes to Environmental Stresses in *Thermus thermophilus* HB8

Abbas Karimi-Fard [1], Abbas Saidi [1*], Masoud Tohidfar [1*], Seyede Noushin Emami [2]

[1] Department of Cell and Molecular Biology, Faculty of Life Sciences and Biotechnology, Shahid Beheshti University, Tehran, Iran
[2] Department of Molecular Biosciences, Wenner-Gren Institute, Stockholm University, SE 106 91 Stockholm, Sweden

Corresponding Author: Abbas Saidi, PhD, Professor, Tel: +989121056599, E-mail: abbas.saidi@gmail.com; Masoud Tohidfar, PhD, Professor, Tel: +989124627014, E-mail: m_tohidfar@sbu.ac.ir; Department of Cell and Molecular Biology, Faculty of Life Sciences and Biotechnology, Shahid Beheshti University, Tehran, Iran.

Received November 26, 2024; **Accepted** March 1, 2025; **Online Published** December 30, 2025

Abstract

Introduction: *Thermus thermophilus* is a thermophilic bacterium known for its resilience in extreme environments. Investigating its transcriptomic responses to environmental stresses can uncover critical adaptive mechanisms.

Materials and Methods: This study analyzed transcriptomic data from 10 microarray datasets, including 63 samples (36 stress-exposed and 27 controls). Stress conditions included copper, cold, zinc, iron, heat, salt, H_2O_2, tetracycline, diamide, and alkylation. Differentially expressed genes (DEGs) were identified through meta-analysis, followed by Gene Ontology (GO) enrichment analysis. Weighted gene co-expression network analysis (WGCNA) was employed to detect stress-associated gene modules. Machine learning approaches—decision tree, logistic regression, random forest, adaptive boosting, SVM-RFE, and XGBoost—were used to prioritize key genes.

Results: Meta-analysis revealed 54 upregulated and 196 downregulated genes under stress. GO analysis highlighted significant enrichment in ion transport, localization processes, and transmembrane transporter activity. WGCNA identified two stress-related modules, cyan and lightcyan. SVM-RFE and XGBoost outperformed other machine learning models with superior accuracy, precision, recall, and F1-scores. *TTHA0798* emerged as a hub gene consistently identified across machine learning and DEG/WGCNA analyses.

Conclusions: This study provides a comprehensive analysis of the stress responses of *T. thermophilus*, identifying *TTHA0798* as a key hub gene. The integration of transcriptomic data, co-expression analysis, and machine learning offers valuable insights into the adaptive mechanisms of this extremophile, paving the way for further functional studies.

Keywords: Bacteria, Environmental Stress, Gene Expression, Meta-Analysis, Machine Learning, WGCNA

Citation: Karimi-Fard A, Saidi A, Tohidfar M, Emami SN. Combining Machine Learning Algorithms with Meta-Analysis and WGCNA to Identify Biomarker-Responsive Genes to Environmental Stresses in *Thermus thermophilus* HB8. J Appl Biotechnol Rep. 2025;12(4):1852-1864. doi:10.30491/jabr.2025.490923.1811

Introduction

Thermus thermophilus, a bacterium known for thriving in extremely harsh environments, has garnered significant scientific interest due to its remarkable resilience.[1] Studying the transcriptomic data of *T. thermophilus* offers a wealth of information on how this microorganism adapts to various environmental stresses.[2] By examining the gene expression profiles under different stress conditions, researchers can uncover the molecular mechanisms and pathways that enable this bacterium to survive and function in such challenging settings. This analysis not only deepens our understanding of *T. thermophilus's* biology but also provides potential insights into biotechnological applications where robust stress responses are essential.

The study of *T. thermophilus* is not only vital for understanding fundamental microbial survival strategies but also holds immense ecological and biotechnological importance. In natural ecosystems, extremophiles like *T. thermophilus* play critical roles in nutrient cycling and maintaining biodiversity in extreme habitats such as geothermal springs and hydrothermal vents. Biotechnologically, the robust stress response mechanisms of *T. thermophilus* have inspired advancements in industrial applications, including the development of thermostable enzymes, stress-resilient biocatalysts, and microbial systems capable of functioning under harsh conditions. These innovations are pivotal in industries such as bioenergy, pharmaceuticals, and environmental remediation. By elucidating the molecular basis of its stress tolerance, this research could provide transformative insights into designing resilient microbial strains and biomolecules with tailored applications, thereby bridging the gap between fundamental biology and practical innovation.

Analyzing transcriptomic data using microarray and RNA-seq technologies is a widely employed approach in bacterial research.[3,4] However, a significant limitation in these analyses include the relatively small number of experiments available, often targeting specific stress conditions rather than encompassing a broad spectrum. To overcome this constraint, researchers utilize meta-analysis, a technique that integrates data from multiple independent datasets. By pooling information from different studies, meta-analysis increases the sample size, thereby enhancing statistical power and enabling a more comprehensive and robust understanding of gene expression patterns across diverse stress conditions.[5]

Identifying differentially expressed genes (DEGs) through meta-analysis offers a broad perspective on gene activity in response to abiotic stress, highlighting which genes are upregulated or downregulated under specific conditions. However, while DEG analysis is essential for understanding individual gene responses, it does not capture the full complexity of gene interactions and regulatory networks. To bridge this gap, Weighted Gene Co-expression Network Analysis (WGCNA) is employed to reveal gene-gene relationships and co-expression patterns that go beyond mere expression levels.[6] By constructing networks of co-expressed genes, WGCNA enables the identification of gene modules associated with particular stress conditions. This integrative approach not only enhances our understanding of the intricate network of gene interactions during stress responses but also uncovers key regulatory genes and pathways, providing deeper insights into the molecular mechanisms governing stress adaptation.

Machine learning (ML) approaches also hold great promise in the analysis of transcriptomic data,[7] particularly for identifying hub genes that play a critical role in stress response pathways. By employing feature selection algorithms, researchers can pinpoint the most influential genes that drive the observed expression patterns under different conditions.[8] This method goes beyond traditional analysis techniques by focusing on the predictive power of each gene, allowing for a more targeted investigation of gene function. Some of the commonly used feature selection algorithms in this context include AdaBoost,[9] decision tree (DT),[10] logestic regression (LR),[11] random forest (RF),[8] SVM-RFE,[12] and XGBoost.[13] These algorithms help prioritize genes based on their contribution to the model's performance, thereby refining our understanding of the molecular mechanisms underlying the stress adaptation in *T. thermophilus* and offering new opportunities for biotechnological innovations.

This research aims to investigate the underlying molecular mechanisms that govern the response to environmental stresses by integrating several widely used analytical methods with ML approaches. This combined strategy is designed to unravel the complex interactions between genes and pathways that enable organisms to adapt to stress conditions. Future research in this field may significantly benefit from these approaches by identifying novel biomarkers linked to abiotic stress tolerance, thereby advancing our understanding of stress adaptation and potentially leading to the development of more resilient crops and biotechnological innovations.

Materials and Methods

The flowchart in figure 1 comprehensively illustrates the research design and all the steps involved in this study.

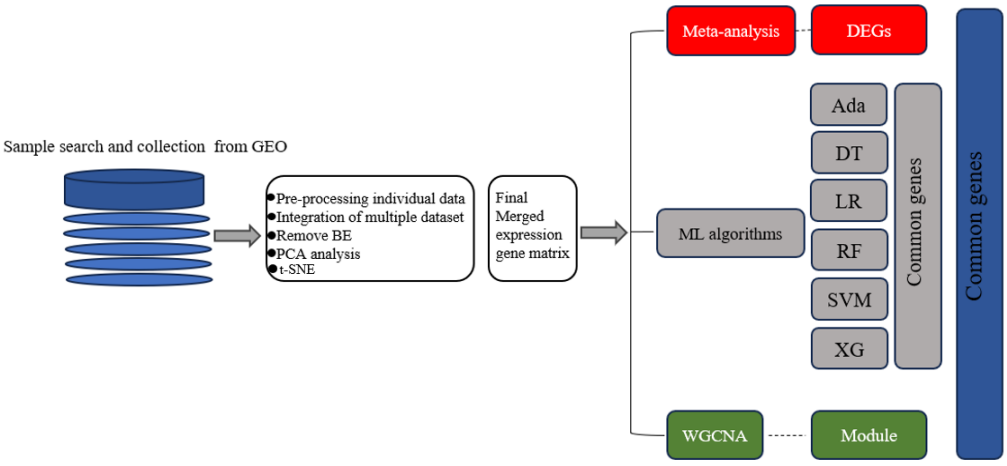

Figure 1. Sequential Workflow for Dataset Integration, Meta-Analysis, WGCNA, and Machine Learning.

Data Collection

This study investigated 10 distinct stress conditions: copper, cold, zinc, iron, heat, salt, hydrogen peroxide (H_2O_2), tetracycline, diamide, and alkylation. For the meta-analysis, a total of 63 samples were selected from the Gene Expression Omnibus (GEO) repository, ensuring a comprehensive

representation of stress responses. The selection process involved identifying studies that specifically examined the aforementioned stress conditions, focusing on those that provided high-quality, well-annotated datasets. From the available studies, 36 stress-exposed samples and 27 control samples were chosen based on criteria such as experimental design, sample size, and relevance to the stress conditions under investigation. This careful selection aimed to minimize biases and enhance the robustness of the meta-analysis, as detailed in Table 1.

Table 1. Details of the Datasets Incorporated in the Research

Accession Number	Sample group Normal: Stress	Type of stress	Platform	Reference
GSE19508	2:1	Copper stress	GPL9209, affymetrix	[15]
GSE19723	2:3	Cold stress	GPL9209, affymetrix	[15]
GSE20900	2:2	Zinc stress	GPL9209, affymetrix	[16]
GSE21199	3:3	Iron stress	GPL9209, affymetrix	[15]
GSE21288	3:3	High temperature	GPL9209, affymetrix	[15]
GSE21289	3:6	Salt stress	GPL9209, affymetrix	[15]
GSE21430	3:9	H_2O_2 stress	GPL9209, affymetrix	[15]
GSE21432	3:3	Tetracycline stress	GPL9209, affymetrix	[15]
GSE21433	3:3	Diamide stress	GPL9209, affymetrix	[15]
GSE27818	3:3	Alkylation stress	GPL9209, affymetrix	[17]

Data Pre-Processing and Check Quality Control

Data pre-processing steps, including normalization, background correction, and probe ID mapping for each dataset, were conducted using the Robust Multichip Average (RMA) algorithm[18] from the "limma" package in R.[19] Subsequently, we merged the normalized expression matrices from each dataset based on 2280 common open reading frames (ORFs).

One of the significant challenges that can compromise our analysis is the presence of batch effects (BE), especially during meta-analysis when integrating multiple datasets. Batch effects occur when samples are derived from different sources or experimental conditions, introducing variability that can mask or confound the true biological signals in the data.[20] To address this issue, the SVA package in R was utilized to remove batch effects among the samples.[21] To validate the effectiveness of batch effect correction, Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) were performed with parameters set to: perplexity 50, learning rate 1000, and silhouette score 0.032. Both PCA and t-SNE together provide a more comprehensive assessment of the data after batch effect correction. PCA is a linear dimensionality reduction technique that captures the most significant variance in the data, allowing us to quickly identify global patterns and trends.[22] However, it may not always reveal complex non-linear relationships between data points. In contrast, t-SNE is a non-linear method that excels at visualizing high-dimensional data by preserving local structures and clustering patterns.[23] By combining these two approaches, we ensure that our analysis captures both the overall trends in the dataset (through PCA) and the finer, more intricate patterns (through t-SNE). This dual approach offers a more robust validation of the effectiveness of batch effect correction, ultimately leading to more reliable interpretations of gene expression data.

After integrating the data and correcting for batch effects, we identified differentially expressed genes (DEGs) using the 'limma' package in R. Genes were deemed significant if they had an absolute log fold change (|logFC|) exceeding 0.5 and an adjusted *p*-value below 0.05.

Gene Ontology (GO) Analysis

Functional enrichment analysis was conducted using the differentially expressed genes (DEGs) to gain insights into their roles in various biological contexts. GO analysis, including categories such as biological processes (BP) and molecular functions (MF), was conducted using the ShinyGO V.0.8 platform.[24] This approach allowed for a detailed examination of the molecular functions, biological pathways, and cellular locations associated with the DEGs.

WGCNA Construction

The WGCNA R package was utilized to construct the co-expression network and identify correlated genes.[6] We filtered the normalized expression matrix to retain genes with an average expression level greater than 10, resulting in the selection of 675 genes from the original 2,280. This approach enhances the signal-to-noise ratio and focuses on biologically relevant genes, improving the reliability of the results and facilitating a more meaningful interpretation of the data. Subsequently, Pearson's correlation coefficient was calculated for pairs of genes. The value of β was established using the scale-free topology index, with an R^2 value greater than 0.8 indicating that the network closely approximates the characteristics of a scale-free network distribution. Subsequently, the similarity matrix was converted into a topological overlap measure (TOM). To detect modules, a hierarchical clustering tree was created, and the cutreeDynamic function was applied with a minimum module size threshold of 10 genes.

The primary functional modules, containing genes with the h ighest correlations, were identified and extracted.

Machine Learning (ML) Algorithms

Due to the complexity of using ML on the initial set of 2280 genes, we first conducted a t-test[25] to filter out genes unlikely to be involved in abiotic tolerance. We set an adjusted *p*-value threshold of 0.0001 for our preliminary selection, which resulted in narrowing down the list to 431 genes.

AdaBoost is an ensemble method that constructs a powerful classifier by integrating several weaker classifiers. During each iteration, it adjusts by placing more emphasis on instances that were incorrectly classified in previous rounds, enhancing the model's accuracy with each step.[26] The AdaBoost model was implemented with these parameters: n_estimators_options = 10, learning_rate = 0.002, and algorithm = 'SAMME.R'.

Decision trees (DT) are a prominent classification method that can be efficiently trained on large datasets, delivering accurate predictions and offering insights into the biological basis behind these predictions, as shown by Clare et al.[27] The DT model was developed using the following parameters: an entropy-based criterion, a maximum depth of 4, features selection determined by the 'log2' method, and a random data splitting strategy.

Logistic Regression (LR) serves as a widely-used statistical technique for binary classification tasks. It estimates the likelihood of a particular input belonging to a specific category by establishing a logistic curve that captures the connection between the dependent and independent variables. The logistic function compresses the output into a range between 0 and 1, allowing for binary classification of inputs.[28] This model was configured with the following parameters: a regularization strength (C) of 0.00001, utilizing the 'sag' solver (Stochastic Average Gradient Descent) which is ideal for large datasets, and applying L2 regularization (penalty = 'l2') to prevent overfitting.

Random Forest (RF) is a versatile machine learning method that constructs numerous decision trees during its training phase, allowing it to efficiently manage both regression and classification problems.[29] It addresses overfitting by averaging the outcomes of these trees, thus enhancing robustness against noisy data. In this study, the RF model was configured with the following parameters: 100 trees (n_estimators = 100), Gini impurity for split quality (criterion = 'gini'), class weights balanced inversely to class frequencies (class_weight = 'balanced'), and a maximum tree depth of 5 (max_depth = 5) to manage model complexity. Support Vector Machine (SVM) is a supervised learning algorithm utilized as outlined in our previous study.[30] XGBoost (Extreme Gradient Boosting) represents an efficient implementation of the gradient boosting algorithm, designed to create a robust predictive model by aggregating the predictions of multiple weak models. Its popularity stems from its exceptional speed and performance in classification tasks, effective handling of missing data, and scalability for large datasets.[31] For this research, XGBoost used logarithmic loss (eval_metric = 'logloss') to evaluate performance. The feature selection was refined using Gain, enhancing both the interpretability and effectiveness of the model. This approach ensured a balance between precision and comprehensibility.

Model Evaluation

In the test set, several performance metrics were estimated, including the Area Under the Curve (AUC), accuracy, precision, recall, and F1 score. The best predictive model was determined based on the highest AUC, reflecting its superior ability to distinguish between classes.[32] This approach ensures the selection of a model with robust predictive performance across various evaluation criteria.

Results

Batch Correction

After integrating 63 samples from 10 different datasets, apparent batch effects (BE) emerged, as illustrated in Figure 2 (a-c). In the boxplot (Figure 2a), the middle lines are inconsistently located, indicating the presence of batch effects. Similarly, in the PCA plot (Figure 2b), samples from specific datasets cluster together, further highlighting BE. Additionally, in the t-SNE plot (Figure 2c), normal and stress samples are mixed, showing the interference of BE.

After correcting for BE, the improvements are evident. The middle lines in the boxplot (Figure 2d) align more consistently, suggesting that BE has been effectively mitigated. In the PCA plot (Figure 2e), normal and stress samples are distinctly categorized into two separate groups, demonstrating clearer differentiation. The t-SNE plot (Figure 2f) also confirms this, with normal and stress samples now well-separated, reflecting the successful elimination of batch effects.

These visualizations underscore the importance of addressing batch effects in multi-dataset analysis to ensure accurate biological interpretations.

Identification of DEGs

The meta-analysis led to the identification of 188 genes, comprising 54 upregulated genes and 134 downregulated genes. Among the downregulated genes, *TTHA0056* had a logFC of -0.97, while among the upregulated genes, *TTHA1359* (Sdrp) had a logFC of 1.6 (Figure 3) (Supplementary file 1).

Functional Analysis of DEGs

The results of the biological process suggest that *T. thermophilus* utilizes several key biological pathways to cope with stress conditions. The most significantly enriched pathways include ion transport, transport, establishment of localization, and localization (Figure 4a). These pathways play crucial roles in maintaining cellular homeostasis and responding to environmental changes. Ion transport and

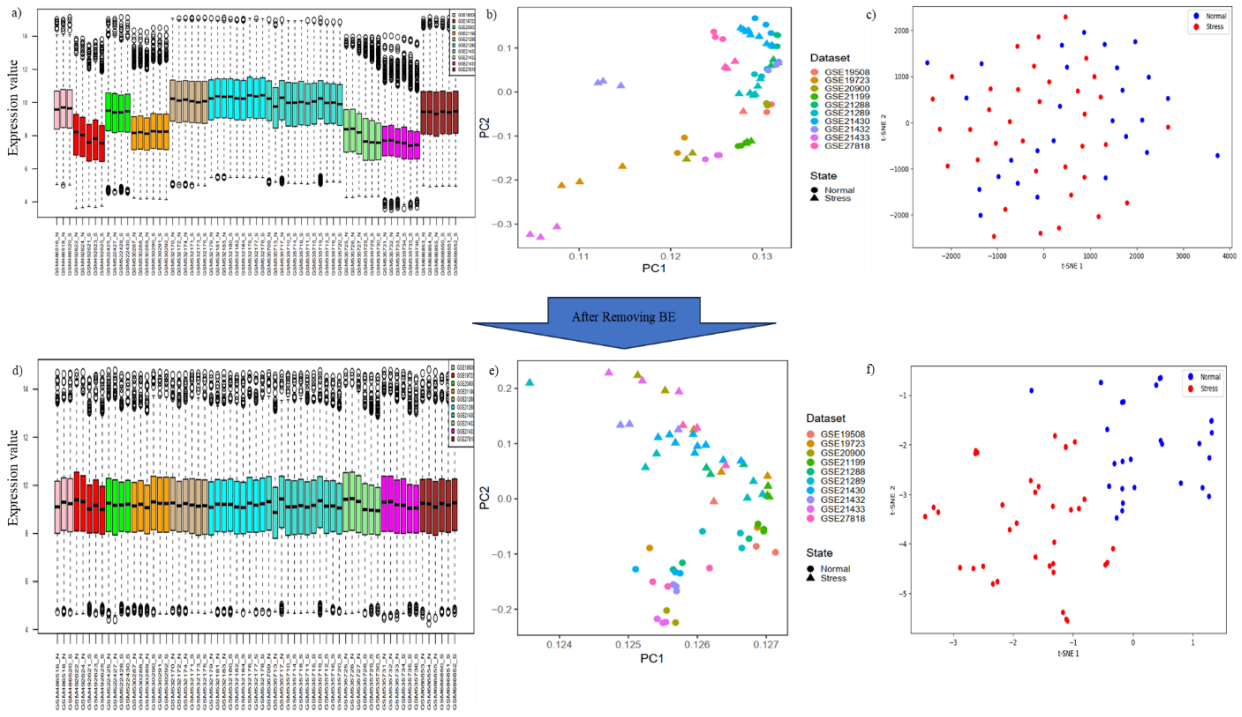

Figure 2. Visualization of Batch Effects (BE) and their Correction. Panels (a-c) show the data before BE correction: (a) Boxplot, (b) PCA, and (c) t-SNE. Panels (d-f) present the data after BE correction: (d) Boxplot, (e) PCA, and (f) t-SNE.

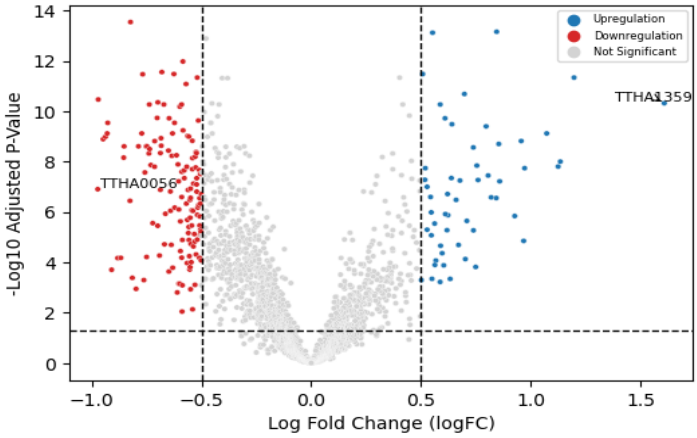

Figure 3. Visualization of DEGs Using a Volcano Plot.

Transport pathways are essential for the movement of ions and molecules across cell membranes, which is critical for balancing osmotic pressure and ensuring the availability of necessary nutrients during stress conditions.[33] The processes involved in "establishment of localization" and "localization" are vital for the precise positioning and distribution of cellular components, allowing the bacterium to adapt its internal structure to external stressors.[34] In the context of stress response in *T. thermophilus*, several significant molecular functions (MF) have been identified, including "transmembrane transporter activity" and "transporter activity" (Figure 4b), which were the most significant terms in the molecular function category. These molecular functions are crucial for the bacterium's ability to adapt to and mitigate the effects of environmental stress.

WGCNA Analysis

Based on the results of the WGCNA, we selected a soft-thresholding power (β) of 12 to construct a scale-free network. This choice of β was made to achieve an optimal balance between scale independence and mean connectivity, as depicted in the plots (Figure 5a).

In this analysis, we identified a total of 13 distinct gene modules, representing groups of highly co-expressed genes.

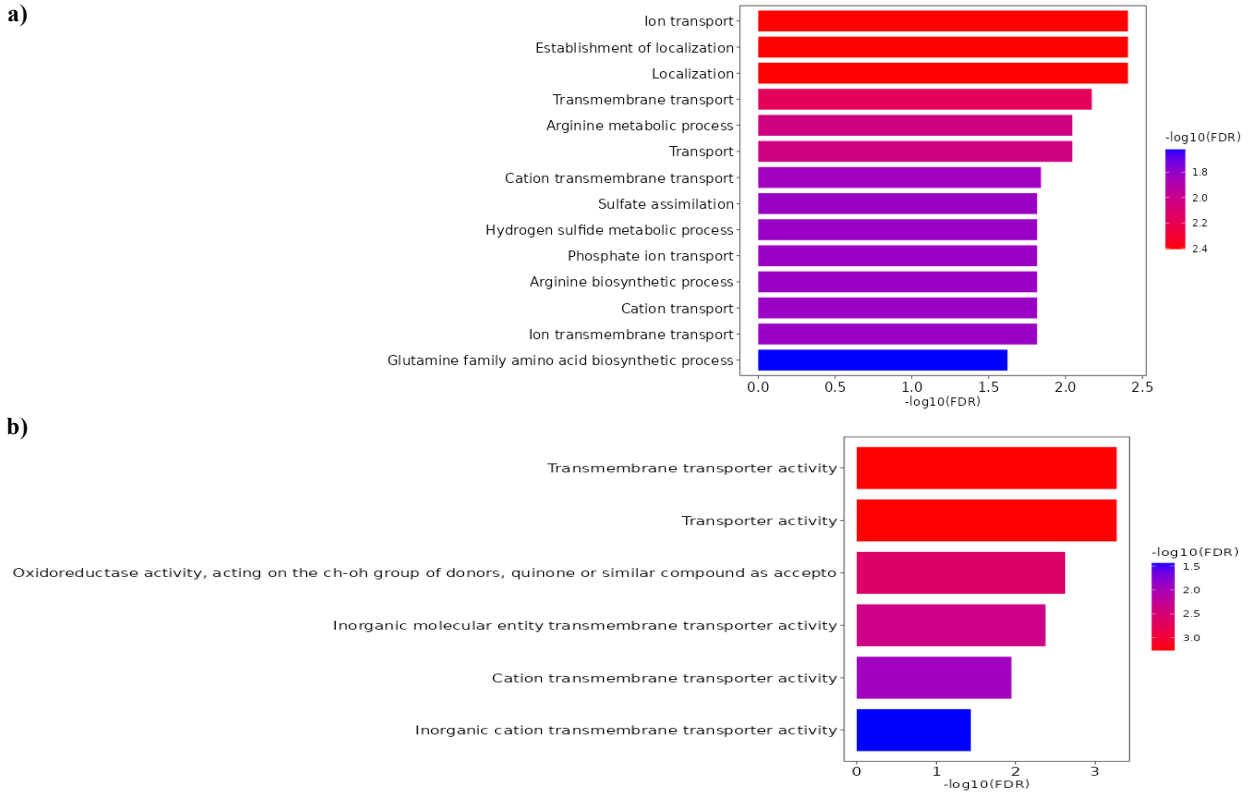

Figure 4. Functional Enrichment Analysis of DEGs: a) Biological Process, and b) Molecular Function.

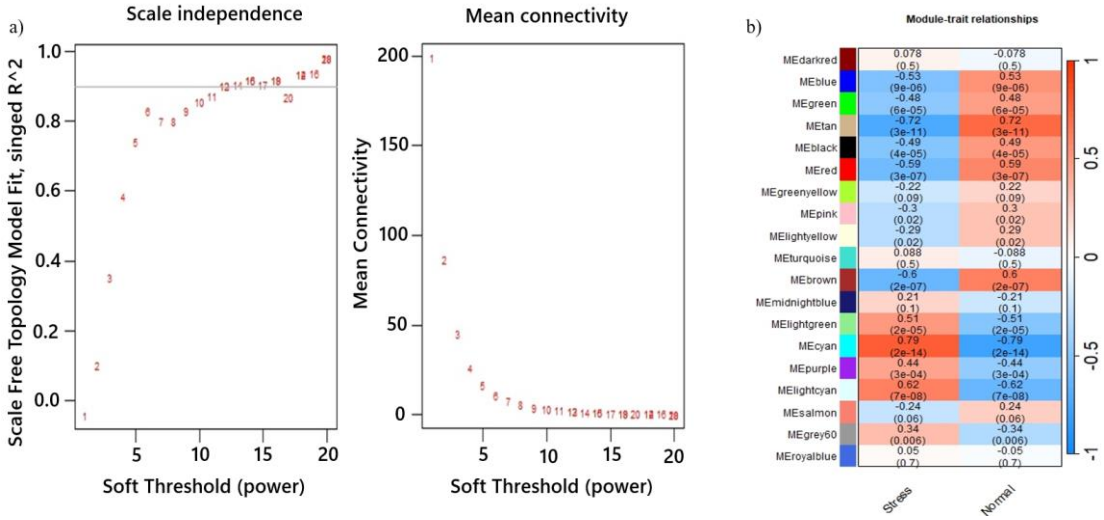

Figure 5. Construction of Co-Expression Modules based on the Integration of 63 Samples. (a) Analyzing network topology across different soft-threshold powers, checking for scale-free topology, with the adjacency matrix defined at a β of 12, b) Module–trait relationships of two traits across 13 modules.

Among these modules, the cyan and lightcyan modules showed the strongest correlation with stress conditions, suggesting their significant involvement in the response to stress. As a result, these two modules were prioritized for further investigation due to their potential roles in biological pathways related to stress adaptation in *T. thermophilus*.

The correlation heatmap between modules and traits (Figure 5b) highlights that the cyan module exhibits a strong positive correlation with stress, with a correlation coefficient of +0.79 and a highly significant *p*-value of 2×10^{-14}. This finding indicates that the genes within the cyan module are upregulated under stress conditions, suggesting their involvement in pathways that are activated during stress responses. Similarly, the lightcyan module also showed a

notable positive correlation with stress, with a correlation coefficient of +0.62 and a significant *p*-value of 7×10^{-8}. Moreover, the cyan module exhibited 22 genes, while the lightcyan module contains 21 genes (Supplementary file 2).

Machine Learning Algorithms for Feature Selection

Six different machine learning algorithms including LR, XGBoost, ADABoost, RF, DT, and SVM were applied to identify the most effective model for distinguishing between stress and normal conditions in *T. thermophilus*. The primary objective of this analysis was to determine the model that provides the best classification performance and to leverage the feature selection capability of these algorithms to identify the most significant features, or genes, that play a crucial role in the bacterium's response to stress.

Performance Evaluation of the Algorithms

The performance metrics for each algorithm, including mean accuracy, precision, recall, and F1-score, are summarized in Table 2. LR achieved a mean accuracy of 0.68, a precision of 0.57, and a recall of 0.95. The F1-score was relatively lower, at 0.73, indicating some imbalance in performance between precision and recall. The XGBoost, on the other hand, performed exceptionally well, with a mean accuracy of 0.94, precision and recall both at 0.97, and an F1-score of 0.95. These values suggest that XGBoost effectively balances all evaluation metrics, making it one of the most robust models for this classification task.

ADABoost showed consistent performance with a mean accuracy of 0.92, precision of 0.93, recall of 0.94, and an F1-score of 0.94. The results indicated a high degree of reliability in distinguishing between stress and normal conditions. Similarly, RF delivered a mean accuracy of 0.94, with both precision and recall at 0.94, and an F1-score of 0.94, demonstrating its strong capability in handling the complexity of the dataset.

The DT model achieved the highest mean accuracy of 0.98, with precision, recall, and F1-score all at 0.94, indicating its effectiveness at correctly classifying stress and normal conditions in this dataset. Finally, the SVM also performed well, with a mean accuracy of 0.97, precision and recall both at 0.92, and an F1-score of 0.94. This suggests a consistent ability to classify the samples accurately, though slightly lower in recall compared to other models like XGBoost.

Table 2. Performance Evaluation of Six Algorithms

Metric	LR	XGBoost	AdaBoost	RF	DT	SVM
Mean Accuracy	0.68	0.937	0.921	0.925	0.921	0.937
Mean Precision	0.571	0.940	0.971	0.942	0.915	0.977
Mean Recall	0.950	0.971	0.889	0.940	0.971	0.914
Mean F1-Score	0.727	0.949	0.927	0.936	0.936	0.939

LR: logistic regression; XGBoost: AdaBoost; RF: random forest; DT: decision tree; SVM: support vector machine.

The effectiveness of these models is assessed using the AUC of the Receiver Operating Characteristic (ROC). This metric provides insight into the model's classification ability by examining the balance between sensitivity (true positive rate) and specificity (false positive rate) at various threshold settings. A model with an AUC value nearing 1 signifies a stronger discriminative capacity between the different classes.

The SVM model stands out with an AUC of 0.97 (Figure 6), highlighting its superior ability to distinguish between stress and normal conditions. This high AUC value signifies that the SVM model is the most effective among the six evaluated algorithms regarding classification performance.

Following closely, the AdaBoost algorithm achieves an AUC of 0.95 (Figure 6). This high AUC score suggests that AdaBoost also performs exceptionally well in distinguishing between classes, making it a highly reliable model for classification tasks. Both LR and XGBoost exhibit an AUC of 0.94 (Figure 6). These values indicate strong classification capabilities, though slightly lower than those of SVM and AdaBoost, demonstrating that they remain effective classifiers.

The RF and DT models both demonstrate an AUC of 0.93 (Figure 6). While still exhibiting solid classification performance, these algorithms are somewhat less effective compared to SVM, AdaBoost, LR, and XGBoost, suggesting room for improvement in their ability to differentiate between stress and normal conditions.

The gene selection frequencies of six machine learning algorithms were evaluated. Each algorithm selected 25 genes, except for XGBoost, which identified 12 genes based on the gain criterion (Figure 7). All models were validated using 5-fold cross-validation with 100 repetitions, ensuring the reliability and robustness of our findings.

The reason XGBoost selected only 12 genes is due to its use of the gain criterion, which identifies genes that significantly improve model accuracy. Gain measures the contribution of each gene to reducing the prediction error; hence, it prioritizes genes with the most impact on the model's performance. In contrast, the RF algorithm used the Gini criterion, which evaluates genes based on their ability to reduce impurity in the data. This approach often results in a broader selection of genes, as it considers a wider range of features that contribute to data classification.

The expression patterns of selected genes, as determined by six different algorithms—AdaBoost (Ada), DT, LR, RF, SVM, and XGBoost exhibit substantial variation in both

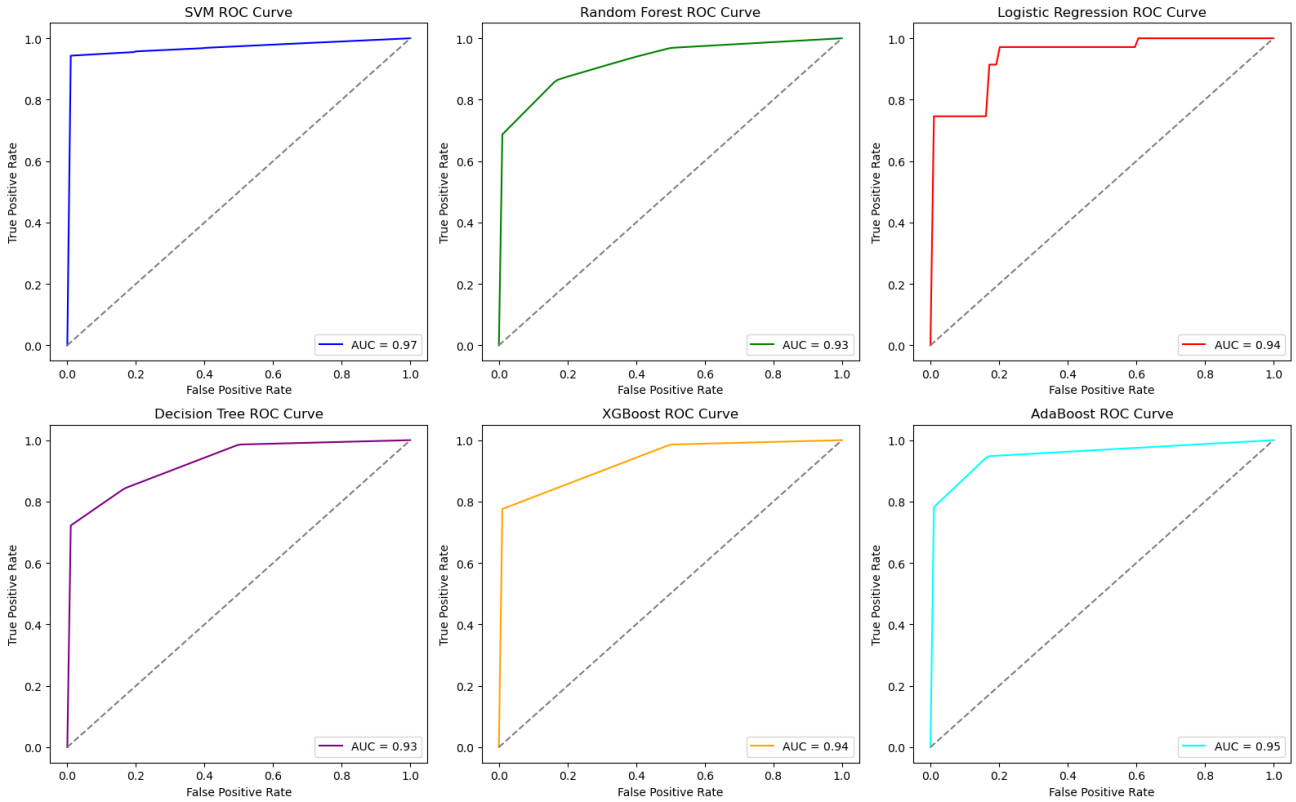

Figure 6. Receiver Operating Characteristic (ROC) Curves and Area Under the Curve (AUC) for Six Classification Algorithms.

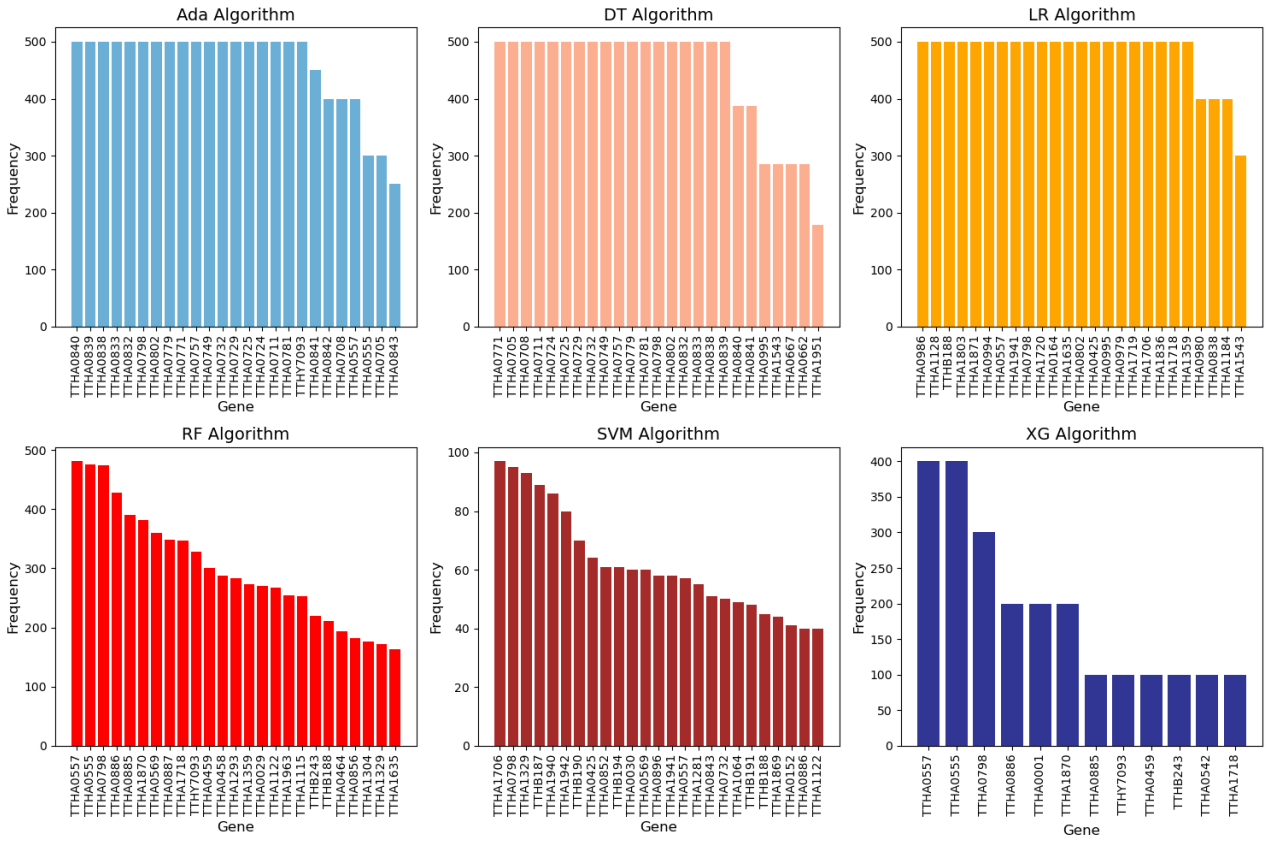

Figure 7. Gene Selection Frequency Across Six Machine Learning Algorithms.

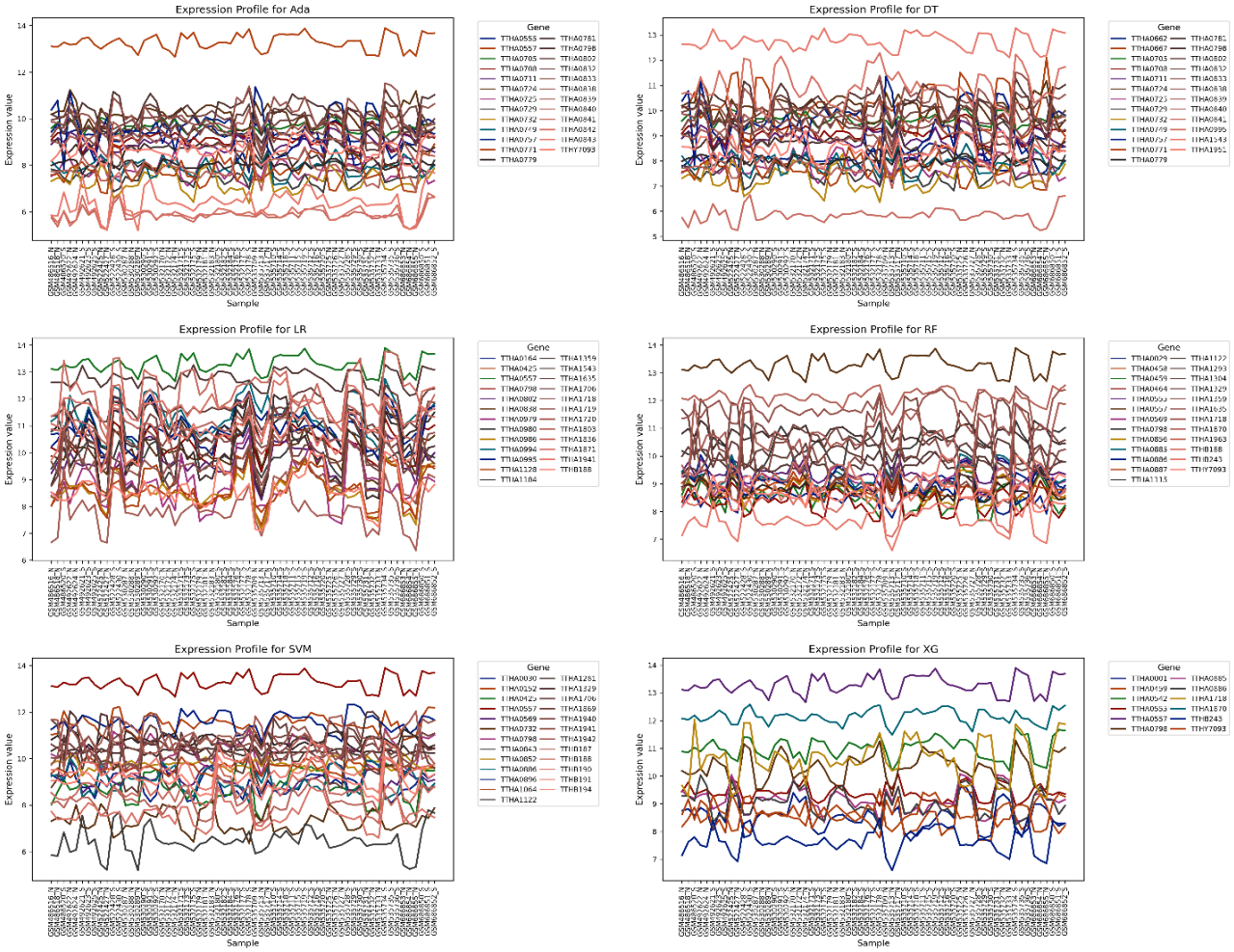

Figure 8. Expression Pattern of 25 Genes Selected by Six Different Algorithms.

intensity and range (Figure 8). Each algorithm appears to prioritize different sets of genes, reflecting their unique methodologies for gene selection. For instance, the Ada algorithm identifies genes with expression levels ranging from 6 to 14, capturing both high and low expression levels. In contrast, the DT algorithm focuses on genes with more moderate expression levels, ranging from 5 to 13, suggesting a potentially more uniform regulatory response under this model. The LR and RF show broader and more diverse expression profiles, with genes such as *TTHA0464* in LR and *TTHA0030* in RF showing clear upregulation across samples. Both algorithms highlight gene expression ranges between 7 and 14, reflecting a preference for higher expression levels. This is similar to the gene selection process seen in SVM, where the expression levels also span from 6 to 14, indicating its tendency to prioritize highly expressed genes akin to Ada and LR. The fluctuation of expression values across all the algorithms suggests a non-linear and complex regulation mechanism, with each model identifying different sets of important genes based on their underlying assumptions.

Gene selection frequencies of six machine learning algorithms were evaluated. Each algorithm selected 25 genes, except for XGBoost, which identified 12 genes based on the gain criterion (Figure 9). All models were validated using a 5-fold cross validation and 100 repetitions, ensuring the reliability and robustness of the findings.

The selected genes from the six machine learning algorithms were compared to identify common genes using a Venn diagram (Figure 9). This comparative analysis allowed for a visual representation of the overlap between the gene sets, facilitating the identification of genes consistently selected by multiple algorithms. This approach allowed us to pinpoint the overlapping genes that are consistently identified as significant across multiple algorithms, thereby highlighting their potential importance in the biological context. The analysis revealed that *TTHA0798* was the only gene commonly selected by all six algorithms, highlighting its pivotal role and robust predictive value across different machine learning techniques. *TTHA0798* belongs to the up-regulated genes with a logFC of 0.84 and is also part of the cyan module. These findings indicate that *TTHA0798* is not only present in the list of differentially expressed genes (DEGs) but is also identified in the WGCNA analysis.

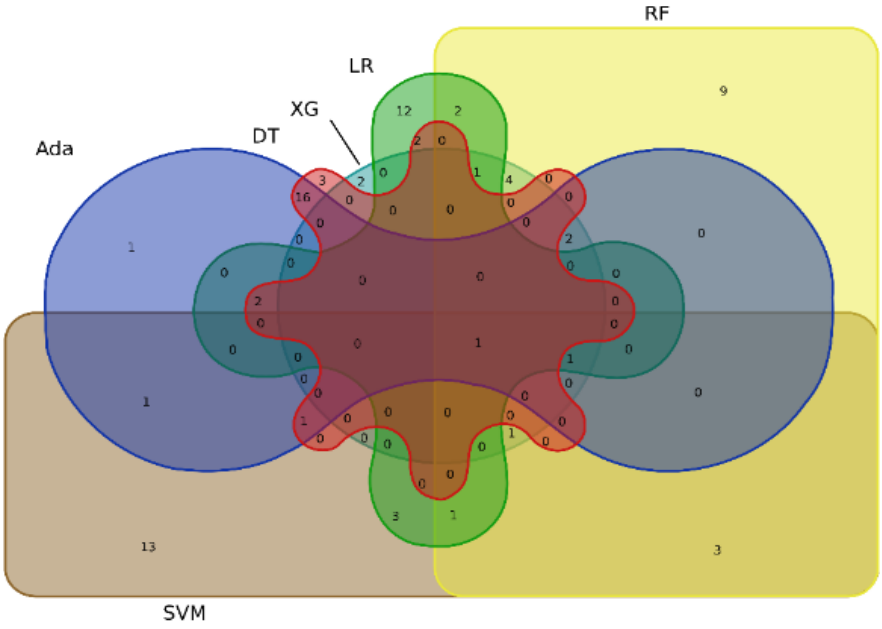

Figure 9. Venn Diagram Analysis of Commonly Selected Genes Across Six Machine Learning Algorithms.

Discussion

In this study, we conducted a comprehensive analysis of the gene expression patterns in *T. thermophilus* under stress conditions using a combination of meta-analysis, WGCNA, and machine learning techniques. Our findings identified a total of 188 differentially expressed genes (DEGs), consisting of 54 upregulated and 134 downregulated genes.

The biological process analysis indicated that *T. thermophilus* employs several key pathways to manage stress conditions effectively. Among these, pathways related to ion transport, transport, establishment of localization, and localization play a crucial role in maintaining cellular homeostasis and adapting to environmental changes. These pathways facilitate the movement of ions and molecules across cell membranes, essential for balancing osmotic pressure and ensuring the availability of nutrients under stress. Moreover, the processes of establishing localization and precise distribution of cellular components enable *T. thermophilus* to reorganize its internal structure in response to external stressors, demonstrating its adaptability. Additionally, significant molecular functions were identified, including oxidoreductase activity, transporter activity, and transmembrane transporter activity, contributing to the bacterium's resilience against environmental stress. These findings collectively highlight the sophisticated biological mechanisms that *T. thermophilus* utilizes to survive and function in extreme environments, emphasizing its potential utility in biotechnology.

The WGCNA revealed crucial insights into the stress response mechanisms of *T. thermophilus*. This analysis identified 13 distinct gene modules, with the cyan and lightcyan modules showing the strongest associations with stress conditions. The robust correlation of these modules with stress implies their significant roles in stress adaptation processes. Specifically, the cyan module, which showed a strong positive correlation with stress, likely contains genes that are upregulated in response to stressors, thus participating in key pathways activated during these conditions. Similarly, the lightcyan module's positive correlation with stress suggests its involvement in biological processes that support the organism's adaptive response to environmental challenges. These findings underscore the relevance of these modules in the stress resilience of *T. thermophilus*, highlighting their potential as targets for further functional studies to understand their contributions to the organism's ability to thrive in extreme environments.

The comparative analysis of six machine learning algorithms, LT, XGBoost, AdaBoost, RF, DT, and SVM provided crucial insights into their effectiveness in distinguishing between stress and normal conditions in *T. thermophilus*. The evaluation of these models using various performance metrics, such as accuracy, precision, recall, and F1-score, revealed that while all models demonstrated competence in classification tasks, XGBoost and SVM stood out for their overall balanced performance. Notably, XGBoost achieved the highest accuracy, precision, and recall scores, making it one of the most robust models for this specific task.

In the context of AUC analysis, SVM exhibited the highest AUC score, underscoring its superior ability to discriminate between stress and normal conditions. The strong performance of AdaBoost also highlights its reliability in handling classification problems. While DT achieved the highest mean accuracy among the models, its slightly lower AUC score suggests that it might be prone to

overfitting the training data, leading to decreased generalizability compared to SVM and XGBoost.

These findings underscore the importance of algorithm selection in bioinformatics research, particularly when the goal is to identify the most significant genes involved in stress adaptation. The success of SVM and AdaBoost in achieving high AUC values indicates their potential to serve as reliable tools for pinpointing critical genes that play pivotal roles in the stress response of *T. thermophilus*. Leveraging the feature selection capabilities of these algorithms could drive future research aimed at understanding the molecular mechanisms underpinning stress resilience, ultimately contributing to biotechnological advancements in fields that benefit from robust microbial adaptation.

The *TTHA0798* has been identified as a hub gene through an integrated approach combining meta-analysis, WGCNA, and machine learning. This gene encodes a protein with a GGDEF domain, which, according to gene ontology analysis, is localized to the plasma membrane and exhibits diguanylate cyclase activity. The primary biological process associated with *TTHA0798* is its involvement in cell adhesion during single-species biofilm formation. Previous report has shown that proteins containing GGDEF domains are involved in the synthesis of cAMP or c-di-GMP[35] and through these molecules, can be involved in signal transduction, especially during stress. The *TTHA1359* (Sdrp) functions as a transcriptional factor, regulating the expression of multiple genes including *TTHA0798*.[15] In our study, *TTHA1359* showed the highest expression levels, with a corresponding increase in the expression of *TTHA0798*, suggesting its regulatory influence. Previous research by Agari et al.[15] indicated that the expression of this protein increases following phage infection. In contrast, our analysis revealed that *TTHA0798* is significantly upregulated under a range of environmental stress conditions, highlighting its broader role in stress response.

While this study provides significant insights into the stress response mechanisms of *T. thermophilus*, it is important to acknowledge its limitations. One key limitation is the identification of hub genes, such as *TTHA0798* and *TTHA1359*, which relies on bioinformatic predictions and necessitates experimental validation to confirm their functional roles in stress adaptation. Additionally, the selection of datasets for meta-analysis may introduce biases, as the available studies might not comprehensively cover all possible stress conditions. Furthermore, the analytical methods employed, including WGCNA and machine learning algorithms, are constrained by the quality and size of the input data. Lastly, while our findings are robust for *T. thermophilus*, their generalizability to other organisms or conditions remains to be explored.

To address these limitations, future research should prioritize the validation of identified differentially expressed genes (DEGs) and gene modules through experimental approaches such as quantitative real-time PCR (qRT-PCR) and gene knockout studies. These validations are essential for confirming the roles of these genes in stress response and for gaining a deeper understanding of the underlying molecular mechanisms. Moreover, investigating the protein-protein interactions and post-translational modifications of the identified hub genes will provide valuable insights into the regulatory networks involved. Exploring the ecological relevance of these genes in the natural habitats of *T. thermophilus* can further illuminate their adaptive significance. Finally, integrating multi-omics data, including proteomics and metabolomics, with transcriptomic analysis will enhance our understanding of the comprehensive stress response mechanisms in *T. thermophilus*, paving the way for potential biotechnological applications.

Conclusion

The integrative approach combining meta-analysis, functional analysis, WGCNA, and machine learning provided valuable insights into the stress adaptation mechanisms of *T. thermophilus*. The identification of *TTHA0798* as a consistently significant gene across multiple analyses highlights its pivotal role in stress response. Our findings suggest that these methodologies can be powerful tools for uncovering the molecular underpinnings of stress adaptation, paving the way for targeted studies on *T. thermophilus* and other extremophiles. Additionally, the use of the SVM model, with the highest AUC and high parameters of precision and F1-score, along with AdaBoost, indicates that these two algorithms are highly effective for feature selection and classification in transcriptomic data of *Thermus thermophilus*.

Authors' Contributions

AKF, AS, and MT performed the literature search and designed, edited, and analyzed the data for the manuscript. They also created all the figures. Contributions to the final manuscript were made by all authors, including AKF, AS, MTF, and SNE. Each author has reviewed and given their approval for the manuscript.

Conflict of Interest Disclosures

The authors declare that they have no conflicts of interest.

References

1. Ohtani N, Tomita M, Itaya M. An extreme thermophile, *Thermus thermophilus*, is a polyploid bacterium. J Bacteriol. 2010;192(20):5499-505. doi:10.1128/jb.00662-10
2. Mega R, Manzoku M, Shinkai A, Nakagawa N, Kuramitsu S, Masui R. Very rapid induction of a cold shock protein by temperature downshift in *Thermus thermophilus*. Biochem Biophys Res Commun. 2010;399(3):336-40. doi:10.1016/j.bbrc.2010.07.065

3. Goswami M, Narayana Rao AV. Transcriptome profiling reveals interplay of multifaceted stress response in Escherichia coli on exposure to glutathione and ciprofloxacin. Msystems. 2018;3(1):10-128. doi:10.1128/msystems.00001-18
4. Jawaharraj K, Peta V, Dhiman SS, Gnimpieba EZ, Gadhamshetty V. Transcriptome-wide marker gene expression analysis of stress-responsive sulfate-reducing bacteria. Sci Rep. 2023;13(1):16181. doi:10.1038/s41598-023-43089-8
5. Vennou KE, Piovani D, Kontou PI, Bonovas S, Bagos PG. Methods for multiple outcome meta-analysis of gene-expression data. MethodsX. 2020;7:100834. doi:10.1016/j.mex.2020.100834
6. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9(1):559. doi:10.1186/1471-2105-9-559
7. Ng S, Masarone S, Watson D, Barnes MR. The benefits and pitfalls of machine learning for biomarker discovery. Cell Tissue Res. 2023;394(1):17-31. doi:10.1007/s00441-023-03816-z
8. Karimi-Fard A, Saidi A, Tohidfar M, Saxena A. Identification of key responsive genes to some abiotic stresses in Arabidopsis thaliana at the seedling stage based on coupling computational biology methods and machine learning. J Appl Biotechnol Rep. 2023;10(3):1079-90. doi:10.30491/jabr.2023.388345.1611
9. Soui M, Mansouri N, Alhamad R, Kessentini M, Ghedira K. NSGA-II as feature selection technique and AdaBoost classifier for COVID-19 prediction using patient's symptoms. Nonlinear Dyn. 2021;106(2):1453-75. doi:10.1007/s11071-021-06504-1
10. Al Snousy MB, El-Deeb HM, Badran K, Al Khlil IA. Suite of decision tree-based classification algorithms on cancer gene expression data. Egypt Inform J. 2011;12(2):73-82. doi:10.1016/j.eij.2011.04.003
11. Huang HH, Liu XY, Liang Y. Feature selection and cancer classification via sparse logistic regression with the hybrid L1/2+ 2 regularization. PloS One. 2016;11(5):e0149675. doi:10.1371/journal.pone.0149675
12. Liang Y, Zhang F, Wang J, Joshi T, Wang Y, Xu D. Prediction of drought-resistant genes in *Arabidopsis thaliana* using SVM-RFE. PLoS One. 2011;6(7):e21750. doi:10.1371/journal.pone.0021750
13. Nguyen DV, Park J, Lee H, Han T, Wu D. Assessing industrial wastewater effluent toxicity using boosting algorithms in machine learning: A case study on ecotoxicity prediction and control strategy development. Environ Pollut. 2024;341:123017. doi:10.1016/j.envpol.2023.123017
14. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. 2012;41(D1):D991-5. doi:10.1093/nar/gks1193
15. Agari Y, Kuramitsu S, Shinkai A. Identification of novel genes regulated by the oxidative stress-responsive transcriptional activator SdrP in *Thermus thermophilus* HB8. FEMS Microbiol Lett. 2010;313(2):127-34. doi:10.1111/j.1574-6968.2010.02133.x
16. Sakamoto K, Agari Y, Agari K, Kuramitsu S, Shinkai A. Structural and functional characterization of the transcriptional repressor CsoR from *Thermus thermophilus* HB8. Microbiology. 2010;156(7):1993-2005. doi:10.1099/mic.0.037382-0
17. Morita R, Hishinuma H, Ohyama H, Mega R, Ohta T, Nakagawa N, et al. An alkyltransferase-like protein from *Thermus thermophilus* HB8 affects the regulation of gene expression in alkylation response. J Biochem. 2011;150(3):327-39. doi:10.1093/jb/mvr052
18. Gautier L, Cope L, Bolstad BM, Irizarry RA. affy—analysis of *Affymetrix GeneChip* data at the probe level. Bioinformatics. 2004;20(3):307-15. doi:10.1093/bioinformatics/btg405
19. Ritchie ME, Phipson B, Wu DI, Hu Y, Law CW, Shi W, et al. *limma* powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47. doi:10.1093/nar/gkv007
20. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010;11(10):733-9. doi:10.1038/nrg2825
21. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012;28(6):882-3. doi:10.1093/bioinformatics/bts034
22. Zhang N, Casasent TD, Casasent AK, Kumar SV, Wakefield C, Broom BM, et al. PCA-Plus: enhanced principal component analysis with illustrative applications to batch effects and their quantitation. bioRxiv. 2024. doi:10.1101/2024.01.02.573793
23. Aliverti E, Tilson JL, Filer DL, Babcock B, Colaneri A, Ocasio J, et al. Projected t-SNE for batch correction. Bioinformatics. 2020;36(11):3522-7. doi:10.1093/bioinformatics/btaa189
24. Ge SX, Jung D, Yao R. ShinyGO: a graphical gene-set enrichment tool for animals and plants. Bioinformatics. 2020;36(8):2628-9. doi:10.1093/bioinformatics/btz931
25. Baldi P, Long AD. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. Bioinformatics. 2001;17(6):509-19.
26. Shalev-Shwartz S, Ben-David S. Understanding machine learning: From theory to algorithms. Cambridge university press. 2014. doi:10.1017/CBO9781107298019
27. Clare A, King RD. Predicting gene function in *Saccharomyces cerevisiae*. Bioinformatics-Oxford. 2003;19(2):42-9.
28. Hosmer Jr, Lemeshow S, Sturdivant RX. Applied logistic regression. John Wiley & Sons. 2013. doi:10.1002/9781118548387
29. Breiman L. Random forests. Machine learning. 2001;45(1):5-32. doi:10.1023/A:1010933404324
30. Karimi-Fard A, Saidi A, TohidFar M, Emami SN. Novel candidate genes for environmental stresses response in *Synechocystis* sp. PCC 6803 revealed by machine learning algorithms. Braz J Microbiol. 2024;55(2):1219-29. doi:10.1007/s42770-024-01338-6
31. Ma B, Meng F, Yan G, Yan H, Chai B, Song F. Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data. Comput Biol Med. 2020;121:103761. doi:10.1016/j.compbiomed.2020.103761
32. Chicco D, Jurman G. The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. BioData Mining. 2023;16(1):4. doi:10.1186/s13040-023-00322-4
33. Joudeh N, Saragliadis A, Schulz C, Voigt A, Almaas E, Linke D. Transcriptomic response analysis of *Escherichia coli* to palladium stress. Front Microbiol. 2021;12:741836. doi:10.3389/fmicb.2021.741836
34. Kannaiah S, Livny J, Amster-Choder O. Spatiotemporal organization of the *E. coli* transcriptome: translation independence and engagement in regulation. Mol Cell. 2019;76(4):574-89. doi:10.17632/jpndmsc3c7.1
35. Ryjenkov DA, Tarutina M, Moskvin OV, Gomelsky M.

Cyclic diguanylate is a ubiquitous signaling molecule in bacteria: insights into biochemistry of the GGDEF protein domain. J Bacteriol. 2005;187(5):1792-8. doi:10.1128/jb.187.5.1792-1798.2005