Original paper

# Prediction of non-coding RNAs in *Fusobacterium nucleatum*-infected mice using machine learning

Pradeep Kumar Yadalam[1,A–D,F], Thilagar Sivasankari[1,A,B,D,F], Muthupandian Saravanan[2,D,F], Kumar Chandan Srivastava[3,D–F], Deepti Shrivastava[1,A–D,F], Maria Maddalena Marrapodi[4,E,F], Marco Cicciù[5,E,F], Giuseppe Minervini[6,E,F]

[1] Department of Periodontics, Saveetha Dental College, Saveetha Institute of Medical and Technical Sciences (SIMATS), Chennai, India
[2] AMR and Nanotherapeutics Lab, Department of Pharmacology, Saveetha Dental College and Hospital, Saveetha Institute of Medical and Technical Sciences (SIMATS), Chennai, India
[3] Department of Oral and Maxillofacial Surgery and Diagnostic Sciences, College of Dentistry, Jouf University, Sakaka, Saudi Arabia
[4] Department of Woman, Child and General and Specialized Surgery, University of Campania Luigi Vanvitelli, Naples, Italy
[5] Department of Biomedical and Surgical and Biomedical Sciences, Catania University, Italy
[6] Multidisciplinary Department of Medical-Surgical and Odontostomatological Specialties, University of Campania Luigi Vanvitelli, Naples, Italy

A – research concept and design; B – collection and/or assembly of data; C – data analysis and interpretation;
D – writing the article; E – critical revision of the article; F – final approval of the article

Dental and Medical Problems, ISSN 1644-387X (print), ISSN 2300-9020 (online) *Dent Med Probl*. 2026;63(1):159–168

**Address for correspondence**
Deepti Shrivastava
E-mail: sdeepti20@gmail.com

**Funding sources**
None declared

**Conflict of interest**
None declared

**Acknowledgements**
None declared

Received on January 30, 2024
Reviewed on April 24, 2024
Accepted on May 25, 2024

Published online on February 27, 2026

**Cite as**
Yadalam PK, Sivasankari T, Saravanan M, et al. Prediction of non-coding RNAs in *Fusobacterium nucleatum*-infected mice using machine learning. *Dent Med Probl.* 2026;63(1):159–168. doi:10.17219/dmp/189304

**DOI**
10.17219/dmp/189304



## Abstract

**Background.** The anaerobic commensal *Fusobacterium nucleatum* is scarce in healthy subgingival dental biofilms but is highly prevalent in periodontal pockets. Numerous genome-wide association studies and gene expression studies using microarrays or RNA sequencing (RNA-Seq) have been performed to better understand the genetic architecture of periodontal disease. However, these investigations have limited predictive capacity for identifying RNAs, particularly non-coding RNAs (ncRNAs). The mechanism of regulation of ncRNAs by *F. nucleatum* to alter disease progression in mice has not been thoroughly investigated.

**Objectives.** The aim of the study was to predict previously uncharacterized ncRNAs in *F. nucleatum*-infected mice using machine learning (ML).

**Material and methods.** Long non-coding RNAs (lncRNAs) and circular RNAs (circRNAs) were identified from the periodontitis gene expression dataset (GSE225589) obtained from the Gene Expression Omnibus (GEO) database and subsequently preprocessed. Long non-coding RNAs and circRNAs were labeled based on the gene expression. Transcriptomic features were analyzed using 3 ML algorithms: random forest (RF); adaptive boosting (AdaBoost); and naïve Bayes (NB). The dataset was labeled and divided into training (80%) and testing (20%) subsets with cross-validation. Additionally, receiver operating characteristic (ROC) curves, confusion matrices and area under the ROC curve (AUC) values were determined.

**Results.** The RF and AdaBoost models outperformed the NB model in classifying lncRNAs and circRNAs. Both RF and AdaBoost achieved an AUC of 100%, whereas the NB model achieved a slightly lower AUC of 92%.

**Conclusions.** This study is the first to apply ML to predict ncRNAs in *F. nucleatum*-infected mice using transcriptomic data. Random forest and AdaBoost showed superior classification performance in identifying lncRNAs and circRNAs associated with the infection. Further studies with larger cohorts and external validation are needed to confirm these findings.

**Keywords:** periodontal disease, machine learning, transcriptomics, non-coding RNAs

## Highlights

- Machine learning enabled the identification of lncRNAs and circRNAs associated with *Fusobacterium nucleatum* infection in mice, marking the first application of machine learning in this context.
- Random forest and AdaBoost achieved perfect classification performance (AUC = 1.000), outperforming naïve Bayes and demonstrating the robustness of ensemble learning approaches for ncRNA prediction.
- The identified ncRNAs demonstrate strong potential as diagnostic or prognostic biomarkers; however, larger datasets and external validation are needed to confirm their clinical applicability.

# Introduction

Chronic multifactorial inflammatory periodontitis progressively destroys the tooth-supporting structures and is associated with dysbiotic dental plaque biofilms.[1,2] Approximately 100 different microbial species inhabit the human oral cavity.[3,4] Although most oral bacteria are commensals, a small proportion are hazardous.[5,6] The formation of bacterial biofilms on tooth surfaces is identified as the main cause of periodontal disease.[7–11] Common oral bacteria include *Porphyromonas gingivalis*, *Tannerella forsythia*, *Prevotella intermedia*, *Campylobacter rectus*, *Eikenella corrodens*, *Fusobacterium nucleatum*, *Aggregatibacter actinomycetemcomitans*, *Treponema* species, and *Eubacterium* species.

Oral bacteria produce various polysaccharides and glycoproteins that enable co-aggregation with planktonic microorganisms and adhesion to surfaces (co-adhesion). The acquired pellicle formed on tooth surfaces, composed primarily of salivary glycoproteins and antibodies, facilitates bacterial attachment. Biofilm-associated microorganisms exhibit distinct characteristics compared with planktonic bacteria, including close spatial organization, production of a self-generated extracellular matrix, reduced metabolic activity, and quorum sensing mechanisms that enhance coordinated survival and persistence.

The most prevalent colonizers are gram-positive facultative anaerobes, particularly *Streptococcus* and *Actinomyces* species.[12] The build-up of dental plaque causes a decrease in oxygen levels, favoring colonization by anaerobic bacteria. *Fusobacterium* species serve as bridging organisms between primary and secondary colonizers.[13] Socransky et al. classified the microorganisms into microbial complexes based on their color.[14] The red and orange complexes are strongly associated with periodontal disease in the subgingival region.[7,15] Among these organisms, *F. nucleatum* is considered a key species in the etiology of periodontitis.

*Fusobacterium nucleatum* is an anaerobic commensal bacterium present in low concentrations in healthy subgingival biofilms[16] but enriched in periodontal pockets.[17] It serves as an important bridging organism between early colonizers and periodontal pathogens.[18] *Fusobacterium nucleatum* has been also associated with colorectal cancer, ulcerative colitis, cardiovascular disease,[19,20] and extraoral infections that might be dangerous in pregnancy,[19] supporting its classification as an opportunistic pathogen.

Experimental periodontitis models in rats have demonstrated that *F. nucleatum* infection can cause abscess formation and alveolar bone loss.[21] Infection with *F. nucleatum* and *A. actinomycetemcomitans* stimulates the production of cytokines such as interleukin (IL)-1. Additionally, tumor necrosis factor (TNF) and IL-17 synergize with IL-1 to enhance the synthesis and expression of additional cytokines (e.g., IL-6), defensins and endothelial activation markers, thereby amplifying the immune response.

Recently discovered non-coding RNA (ncRNAs) include long non-coding RNAs (lncRNAs) and circular RNAs (circRNAs).[22,23] Many diseases are initiated and progress through cis- and trans-regulatory gene expression mechanisms.[24–26] LncRNAs affect the gene regulatory network of host–pathogen interactions and influence biological processes such as cell proliferation, motility, and immune or inflammatory response in cardiovascular, inflammatory and autoimmune diseases.[27–29] The biological functions of circRNAs are not yet fully understood. According to previous studies, circRNAs may function as microRNA (miRNA) sponges, thereby inhibiting miRNA activity, increasing the expression of miRNA target genes, modulating cytokine expression, promoting cell cycle progression, and inhibiting apoptosis.[30–32] Both lncRNAs and circRNAs have been identified as potential diagnostic markers.[33,34]

However, in vitro and in vivo studies have investigated *F. nucleatum* invasion and host response mechanisms.[35–39] Previous research has shown that *F. nucleatum* infection activates the inflammatory response, induces cytokine production and sends danger signals to human glomerular endothelial cells (GECs).[37,40] The molecular mechanisms underlying immune responses to *F. nucleatum* infection in mice are unknown. To understand the genetic architecture of this complex trait, several genome-wide association studies and gene expression investigations using microarrays or RNA sequencing (RNA-Seq) have

been conducted. The outcomes of these investigations, however, do not provide robust insights. Li and Liang developed LncDC, a Python-based tool that outperformed 6 other algorithms in classifying lncRNAs and mRNAs.[41] Using osteosarcoma transcriptomic data, they identified 97 novel lncRNAs, providing potential diagnostic biomarkers or therapy targets.[41] However, no studies have investigated how *F. nucleatum* regulates ncRNAs to influence disease progression in mice. The application of alternative methodologies can improve the generalizability of the results. Machine learning (ML) techniques, combined with resampling strategies, offer a powerful framework for identifying ncRNAs in mice infected with *F. nucleatum*.

Due to their computational effectiveness in identifying generalizable patterns from high-dimensional datasets generated from small samples, ML techniques have been utilized to analyze high-throughput deep sequencing data.[42] Therefore, the current study aimed to apply ML to predict previously uncharacterized ncRNAs in mice infected with *F. nucleatum*.

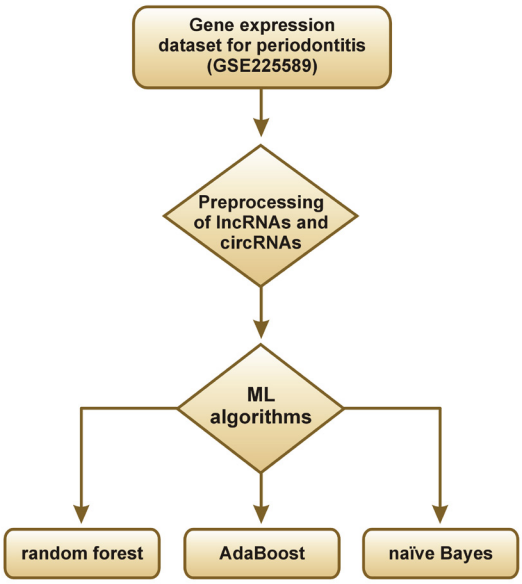


Fig. 1. Flowchart of research methodology
ML – machine learning; AdaBoost – adaptive boosting.

# Material and methods

## Source and processing of data

The gene expression dataset for periodontitis (GSE225589) was retrieved from the Gene Expression Omnibus (GEO) database. A comprehensive search of microarray studies was conducted using the keywords "periodontitis" and "*Rattus norvegicus*". The epitranscriptomic gene expression data was exported to Microsoft Excel (Microsoft Corp., Redmond, USA), and outliers were removed. The dataset was preprocessed to identify lncRNAs and circRNAs, which were then classified and labeled based on their gene expression features. The processed data was subjected to exploratory data analysis.

The Orange Data Mining tool (https://orangedatamining.com/download) was used for gene expression analysis, including data upload, transformation, user-interactive visualization, as well as model inference and visualization. These stages involve data input and transformation, interactive visualization, drawing conclusions about data models, and model representation. Data is typically received, processed and visualized by a workflow component, which then generates the analysis for further processing (Fig. 1,2).

Random forest (RF) parameters included up to 100 decision trees, with no strict limitation on tree depth and feature selection. Adaptive boosting (AdaBoost) parameters included the number of boosting iterations and learning rate with several estimators set to 50, and the SAMME (Stagewise Additive Modeling using a Multi-class Exponential loss) learning algorithm. Data balancing techniques, such as oversampling, undersampling and class weighting, were used to address class imbalances in the dataset. Model performance was evaluated using a confusion matrix, receiver operating characteristic (ROC) curve, and area under the ROC curve (AUC). The AUC provides an aggregate performance measure across different classification thresholds, while the confusion matrix summarizes information about the model's performance in each class. The ROC curve provides a trade-off between sensitivity and specificity, thus facilitating the selection of an appropriate threshold.

Genes with constant expression patterns within the same class were assigned greater weight than genes with erratic expression. This method reduces non-informative genes to improve classification accuracy.

## Adaptive boosting

Adaptive boosting is an ensemble learning algorithm that combines weak learners to build a strong classifier. The algorithm fits a classifier to the original dataset and then trains additional classifiers on weighted versions of the data. In each iteration, the weights of misclassified instances are increased, allowing the model to focus on more challenging cases. AdaBoost can be applied to gene expression data from DNA microarray and RNA-Seq platforms for classification tasks.

## Random forest

Random forest is a sophisticated ML method used for classification and regression. It constructs numerous decision trees during the training phase. The output of this process is the class, that is, the mode of the classes (classification) or the mean prediction (regression) of the individual trees. Random forest can predict clinical

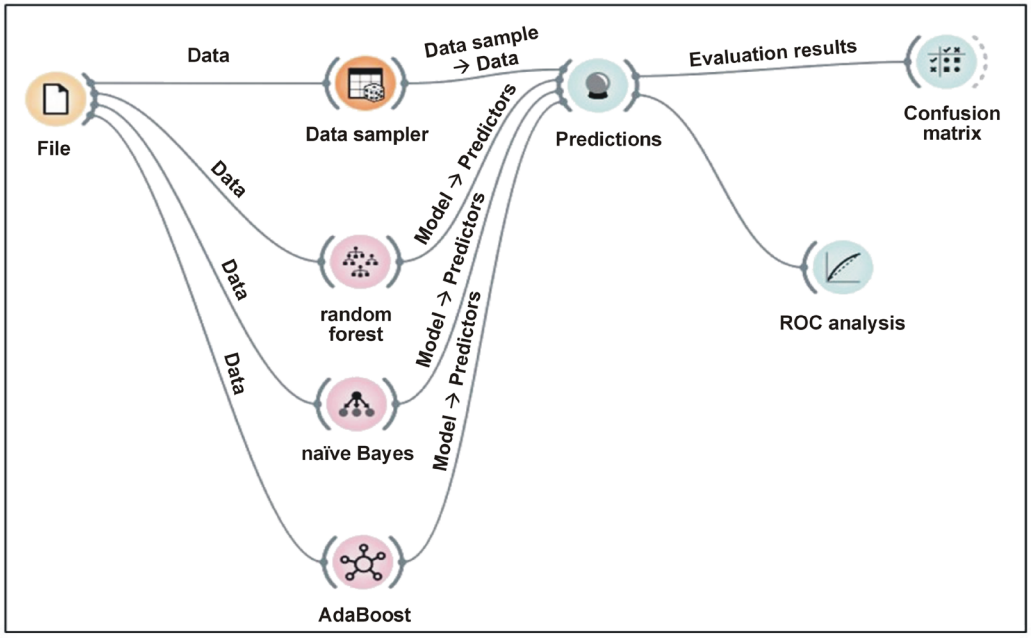


Fig. 2. Orange machine learning (ML) framework
ROC – receiver operating characteristic.

outcomes in gene expression (genes) using profile datasets with several characteristics. The prediction task is a classification problem involving feature representations. The primary objective of this approach is to achieve high classification accuracy. Random forest is effective for the analysis of gene expression datasets, due to its capacity to manage many input variables and their complex interactions. High-dimensional data, such as gene expression profiles, is prone to overfitting; however, this method is robust.

## Naïve Bayes

Naïve Bayes (NB) is a basic yet successful ML method that applies Bayes' theorem with strong (naïve) independence assumptions among features. Due to its simplicity and efficiency, NB is frequently applied to high-dimensional datasets, including gene expression data.

The naïve Bayes classifier (NBC) is widely used in pattern recognition tasks involving gene expression. However, classical estimations of location and scale parameters are sensitive to outliers, which complicates the interpretation of gene expression data using the classical NBC.

# Results

The predictive performance of lncRNA and circRNA disease associations was evaluated using stratified 20-fold cross-validation. In each iteration, the dataset was partitioned into training and testing subsets, with most samples used for training and the remainder for testing. During each cycle, similarities between lncRNAs and circRNAs were recalculated based solely on known training associations.

After estimating the association probabilities of the test samples, the samples were arranged according to their association scores. Samples with higher scores were considered more likely to represent true lncRNA–circRNA disease associations. A sample was deemed positive if an observed association existed in the lncRNA–cirRNA disease node pair and its association score exceeded a predefined threshold. The true positive rate (TPR) and the false positive rate (FPR) were calculated as follows (Equation 1 and Equation 2):

$$TPR = TP/(TP + FN) \quad (1)$$

$$FPR = FP/(FP + TN) \quad (2)$$

where:
FN – number of false negatives (incorrectly classified positive samples);
FP – number of false positives (incorrectly classified negative samples);
TP – number of true positives (correctly classified positive samples);
TN – number of true negatives (correctly classified negative samples).

By varying the threshold, ROC curves were generated. The AUC provided an overall measure of the predictive capability of the model.[43]

Due to a considerable imbalance between observed lncRNA–disease associations (positive samples) and unobserved associations (negative samples), precision–recall (PR) curves and the area under the PR curve (AUPR)

were used to assess predictive performance.[44] Precision and recall were defined as follows (Equations 3 and 4):

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \quad (3)$$

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}) \quad (4)$$

A precision of 1 indicates that FP = 0. Similarly, a recall of 1 indicates that FN = 0. An ideal classifier achieves precision and recall of 1, corresponding to 0 false positives and false negatives. As the number of false negatives increases, the recall decreases because the denominator (TP + FN) grows relative to TP.

To jointly evaluate precision and recall, the F1 score was calculated (Equation 5):

$$\text{F1 score} = \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \times 2 \quad (5)$$

The AUC–ROC curves and, most importantly, recall, precision, specificity, and accuracy, were derived from the confusion matrix. The target feature in this test consisted of 2 groups of predictors: lncRNAs and circRNAs. In this study, stratified 20-fold cross-validation was applied. The confusion matrix summarizes the number of true positives, true negatives, false positives, and false negatives generated by the model on the test data.

## Identification and prediction of non-coding RNAs and construction of classification models

Twenty-fold cross-validation was used to determine model accuracy and AUC after classification using the ML framework (Fig. 2). The outcomes demonstrated that, for lncRNAs and circRNAs, accuracy and AUC value reached 100%. To classify lncRNAs and circRNAs, RF, AdaBoost and NB classification models were built. The RF and AdaBoost models achieved the highest accuracy (AUC = 1.000), whereas the NB model achieved a slightly lower AUC of 0.926 (Table 1). These findings indicate that ML-based classification algorithms targeting lncRNAs and circRNAs demonstrate high diagnostic accuracy.

## Prediction and expression of non-coding RNAs

The ROC analysis was conducted to assess the predictive utility of the selected lncRNAs and circRNAs. All targeted ncRNAs demonstrated AUC values greater than 0.7, indicating acceptable predictive performance. These findings suggest that the identified ncRNAs may serve as prognostic markers in mice infected with *F. nucleatum*. However, the individual AUCs of the selected lncRNAs and circRNAs were lower than the overall classification performance of the RF, AdaBoost and NB models, suggesting that the integrated ML approach provides superior predictive accuracy compared with single biomarkers.

Evaluation of the results using the confusion matrix of the RF model correctly classified 110 lncRNAs as true positives as 114 circRNAs as true negatives (Table 2).

Similar to the RF model, AdaBoost correctly identified 110 lncRNAs as true positives and 114 circRNAs as true negatives (Table 3).

The NB model correctly classified 98 lncRNAs as true positives and 98 circRNAs as true negatives. However, it misclassified 12 lncRNAs as false negatives and 16 circRNAs as false positives (Table 4).

**Table 2.** Confusion matrix of the random forest (RF) model

| Non-coding RNAs | | Actual | | |
|---|---|---|---|---|
| | | lncRNAs | circRNAs | ∑ |
| Predicted | lncRNAs | 110 | 0 | 110 |
| | circRNAs | 0 | 114 | 114 |
| | ∑ | 110 | 114 | 224 |

The table presents the predicted and actual classification labels generated by the model. LncRNAs and circRNAs represent the 2 predicted classes. The data is presented as the number of samples in each category. The bottom row and the rightmost column indicate the total number of samples per class and the overall sum of samples, respectively.

**Table 3.** Confusion matrix of the adaptive boosting (AdaBoost) model

| Non-coding RNAs | | Actual | | |
|---|---|---|---|---|
| | | lncRNAs | circRNAs | ∑ |
| Predicted | lncRNAs | 110 | 0 | 110 |
| | circRNAs | 0 | 114 | 114 |
| | ∑ | 110 | 114 | 224 |

The table presents the predicted and actual classification labels generated by the model. LncRNAs and circRNAs represent the 2 predicted classes. The data is presented as the number of samples in each category. The bottom row and the rightmost column indicate the total number of samples per class and the overall sum of samples, respectively.

**Table 1.** Performance metrics of the machine learning (ML) algorithms used for classifying long non-coding RNAs (lncRNAs) and circular RNAs (circRNAs)

| Model | AUC | CA | F1 score | Precision | Recall | MCC |
|---|---|---|---|---|---|---|
| Random forest | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| AdaBoost | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| Naïve Bayes | 0.926 | 0.875 | 0.875 | 0.876 | 0.875 | 0.751 |

AdaBoost – adaptive boosting; AUC – area under the receiver operating characteristic (ROC) curve; CA – classification accuracy; MCC – Matthews correlation coefficient.

Table 4. Confusion matrix of the naïve Bayes (NB) model

| Non-coding RNAs | | Actual | | |
|---|---|---|---|---|
| | | lncRNAs | circRNAs | ∑ |
| Predicted | lncRNAs | 98 | 12 | 110 |
| | circRNAs | 16 | 98 | 114 |
| | ∑ | 110 | 114 | 224 |

The table presents the predicted and actual classification labels generated by the model. LncRNAs and circRNAs represent the 2 predicted classes. The data is presented as the number of samples in each category. The bottom row and the rightmost column indicate the total number of samples per class and the overall sum of samples, respectively.

## Discussion

*Fusobacterium nucleatum* is a gram-negative, obligate anaerobic bacillus named for its slender, spindle-shaped morphology.[45] It has been associated with the etiology of periodontitis and contributes to dental plaque formation.[46] Acting as a bridging organism between commensal bacteria and periodontal pathogens on tooth and epithelial surfaces, *F. nucleatum* plays a crucial role in mediating physical interactions between gram-positive and gram-negative bacteria.[47] The bacterium expresses multiple adhesins that facilitate binding to other microorganisms and cells, thereby enhancing pathogenicity. The most important virulence factor in *F. nucleatum* is *Fusobacterium* adhesin A (FadA), an adhesion protein.[19] *Fusobacterium* adhesin A exists as mature FadA (mFadA), a secreted protein of 111 amino acids, and pre-FadA, a 129-amino-acid precursor form.[48] The active complex (FadAc), composed of both forms, enables binding and invasion of host cells.[48,49]

Machine learning applications in transcriptomics have quickly expanded, enabling computational analysis of gene expression data generated by techniques such as RNA-Seq. Using ML approaches, researchers can detect differentially expressed genes, categorize samples into groups, predict gene functions, and find hidden molecular patterns. These capabilities provide novel biological insights and support biomarker discovery. A variety of ML methods, including decision trees, support vector machines, RFs, and deep learning models, have been applied to address the high dimensionality and complexity of transcriptomic datasets. Such strategies have great potential to contribute to our understanding of disease diagnosis, gene regulation, and the development of personalized medicine.[50] Because experimental biological research is often time-consuming and costly, computational prediction of disease-associated lncRNAs via bioinformatics has become increasingly common. In recent years, many lncRNA–disease association prediction (LDAP) models have been proposed, including models based on biological networks, models independent of known lncRNA–disease relationships, and ML-based frameworks.

Non-coding RNAs are RNA molecules that do not encode proteins but play significant regulatory functions in several biological processes. Once considered transcriptional noise, ncRNAs are now recognized due to their numerous applications and influence on gene expression.

Well-established subclasses of RNAs include transfer RNAs (tRNAs), which are essential for protein synthesis by transporting amino acids to the ribosomes, and ribosomal RNAs (rRNAs), which form the structural core of ribosomes involved in protein synthesis. Extensive research has been conducted on these RNAs due to their pivotal role in the fundamental functioning of cells. In addition, messenger RNAs (mRNAs) can bind to short RNA molecules, called miRNAs, which can either promote or hinder the translation of mRNAs. MicroRNAs regulate gene expression and play a role in a variety of cellular functions, such as differentiation, development and disease progression.

Another well-known family of ncRNAs is lncRNAs. These larger RNA molecules are transcribed from the genome, yet they do not encode proteins. LncRNAs are involved in several regulatory processes, including post-transcriptional processing, chromatin remodeling and transcriptional control. Additionally, they have been linked to crucial biological processes such as cellular differentiation, the etiology of disease and embryonic development.

Recent studies have also highlighted the functions of enhancer RNAs (eRNAs) and circRNAs in gene regulation. CircRNAs are covalently closed RNA molecules produced during transcription by a back-splicing process. These cells can affect the expression of genes through interactions with RNA-binding proteins or by acting as miRNA sponges. While eRNAs are expected to facilitate enhancer–promoter interactions and modulate gene transcription, they are translated from enhancer regions of the genome.

The functional roles and regulatory mechanisms of ncRNAs are the subject of ongoing research. Dysregulation of ncRNAs has been associated with numerous diseases, including cancer, neurological disorders, periodontal disease, and cardiovascular conditions. The potential of ncRNAs in diagnostics and therapeutics could be utilized in the development of novel biomarkers and focused therapies.

In the present study, we aimed to predict ncRNAs associated with *F. nucleatum* infection in mice to better understand the etiology of *F. nucleatum*-related disorders. Orange, an open-source data visualization and ML toolkit developed in Python, has been used to build and analyze predictive models. The tool supports popular algorithms like RF, neural networks, AdaBoost, NB, and logistic regression, but in our study, 3 models were used: RF; AdaBoost; and NB.

The gene expression dataset for periodontitis (GSE225589) was obtained from the GEO database.

Using GEO2R, 250 differentially expressed genes (DEGs) were identified by comparison with appropriate controls. The 3 ML algorithms, namely RF, AdaBoost and NB, were evaluated for accuracy. The RF and AdaBoost models achieved superior and consistent accuracy compared with the NB model. Specifically, RF achieved an AUC of 1.000, classification accuracy (CA) of 1.000, F1 score of 1.000, precision of 1.000, recall of 1.000, and Matthews correlation coefficient (MCC) of 1.000. AdaBoost demonstrated identical performance metrics. In contrast, the NB model achieved an AUC value of 0.926, CA of 0.875, F1 score of 0.875, precision of 0.876, recall of 0.875, and MCC of 0.751.

Integrating proteomics or metabolomics data could further enhance understanding of disease mechanisms and improve predictive performance. Machine learning-based identification of indicators or pathways of periodontitis may facilitate the development of targeted therapeutic strategies or diagnostic tools.[51]

Several limitations of the present study should be acknowledged, including the relatively small sample size, reliance on a single dataset, and the lack of external validation. The identified DEGs and predictive models require replication in independent datasets to confirm accuracy. Understanding the biological significance and mechanisms of these genes is crucial for further insights. Future research should incorporate multi-omics integration, longitudinal analyses, functional experiments, assessment of potential confounding factors, and development of clinically applicable predictive models. These approaches could provide a more comprehensive understanding of the molecular mechanisms underlying periodontitis, identify novel biomarkers, examine gene expression changes over time, validate the biological relevance of DEGs, and improve patient management and outcomes.

A previous study investigating oral colonization of mice with *P. gingivalis*, *Treponema denticola* and *T. forsythia* demonstrated enhanced intrabony defects and alveolar bone resorption (ABR) during polymicrobial infection.[47] These pathogens successfully established oral colonization and induced ABR. Another research demonstrated that chronic oral infection with *F. nucleatum* has been shown to induce symptoms of periodontal disease in mice, spread via hematogenous routes, alter the host immune system and periodontal risk factors, and cause both pro- and anti-inflammatory reactions.[49] According to earlier in vitro studies, the significantly elevated levels of immunoglobulin G (IgG) and IgM after chronic infection suggest that *F. nucleatum* functions as a powerful B-cell mitogen.[52] Additionally, the mitogenic activity has been attributed to the Toll-like receptor 2 (TLR2) adjuvant outer membrane porin FomA.[53]

The endogenous retroviral-associated adenocarcinoma lncRNA (EVADR) and keratin-7 antisense RNA (KRT7-AS) were among 43 upregulated lncRNAs identified in a recent transcriptome investigation of *F. nucleatum*-infected colon cancer cells.[54] *Fusobacterium nucleatum* has been reported to promote the progression of oral squamous cell carcinoma (OSCC).[54] A lncRNA, MIR4435-2HG-5p, has been shown to be upregulated in *F. nucleatum*-infected OSCC cells, where it functions as a miRNA-296-5p sponge, activates AKT2 signaling, and contributes to AKT2-induced carcinogenesis.[55–59]

The role of lncRNAs in *F. nucleatum*-infected mice sheds light on the complex regulatory mechanisms underlying diseases associated with this pathogen. Previous studies suggest that *F. nucleatum* infection can alter host lncRNA expression, which may be crucial in regulating host immune responses, promoting bacterial survival and accelerating disease progression. One of the main functions of lncRNAs in infected mice is the control of host immune responses. LncRNAs can function as molecular scaffolds or ploys, interacting with proteins or other RNA molecules to regulate signaling pathways and gene expression involved in immune control. Additionally, they can control the expression of pro- and anti-inflammatory genes, ultimately affecting the host's ability to mount an effective immune response against *F. nucleatum*.

A recent study used mono- and dinucleotide sequences to predict ncRNA regulatory functions.[60] A back propagation (BP) neural network with principal component analysis and the Levenberg–Marquardt algorithm trained ncRNAs with accuracies of 81.3% for mixed bacterial ncRNAs and 93.3% for prokaryotic tRNAs.[60] Another study trained a classifier (MncR) using RNAcentral data and reported over 97% accuracy in classifying ncRNA classes.[61] More recently, ML methods such as logistic regression, RF, eXtreme Gradient Boosting (XGBoost), and decision trees have been employed to distinguish coding from non-coding transcripts and classify ncRNAs. In human datasets, RF achieved accuracies exceeding 83%. Our study obtained similar and highly accurate results.[51]

Additionally, lncRNAs may influence cellular functions such as cell division, apoptosis and epithelial–mesenchymal transition, all of which are significant in *F. nucleatum*-associated disorders. Dysregulation of specific lncRNAs in infected mice may promote aberrant cell behavior, potentially contributing to tumor growth, invasion or metastasis in malignancies associated with *F. nucleatum*. Furthermore, lncRNAs may participate in the direct communication between host cells and *F. nucleatum*. According to recent research, bacterial lncRNAs may be transferred to host cells, where they can modulate the expression of host genes and signaling pathways. These lncRNAs may play a role in the establishment and persistence of *F. nucleatum* infection by interfering with host cellular functions.

To fully elucidate the predicted roles of lncRNAs in *F. nucleatum*-infected mice, further studies using experimental and computational methods are necessary. Integrating transcriptomic analysis with functional assays,

including knockdown or overexpression studies, will help clarify the involvement of particular lncRNAs to immune modulation, bacterial persistence and disease development. Additionally, integrative methodologies, such as network analysis and ML, can facilitate the identification of key lncRNA–mRNA interaction networks and improve understanding of the complex regulatory networks.

Investigating lncRNA-mediated regulatory mechanisms in *F. nucleatum* infection enhances our understanding of host–pathogen interactions and identifies potential therapeutic targets aimed at regulating immune responses, limiting bacterial survival or delaying the onset of disease in *F. nucleatum*-related conditions.

# Conclusions

This study represents the first application of ML approaches to predict ncRNAs in a mouse model infected with *F. nucleatum*. The superior performance of the RF and AdaBoost models highlights the robustness of ensemble learning for transcriptomic classification. The identified lncRNAs and circRNAs may serve as promising candidates for biomarker development and for advancing our understanding of host–pathogen regulatory mechanisms. Nevertheless, validation in larger and independent datasets, along with functional experimental studies, is essential to confirm their biological and clinical relevance.

## Ethics approval and consent to participate

Not applicable.

## Data availability

The datasets generated and/or analyzed during the current study are available from the corresponding author on reasonable request.

## Consent for publication

Not applicable.

## Use of AI and AI-assisted technologies

Not applicable.

### ORCID iDs

Pradeep Kumar Yadalam https://orcid.org/0000-0002-6653-4123
Thilagar Sivasankari https://orcid.org/0000-0001-5420-5253
Muthupandian Saravanan https://orcid.org/0000-0002-1480-3555
Kumar Chandan Srivastava https://orcid.org/0000-0002-5969-6810
Deepti Shrivastava https://orcid.org/0000-0002-1073-9920
Maria Maddalena Marrapodi https://orcid.org/0000-0002-9494-6942
Marco Cicciù https://orcid.org/0000-0003-2311-9728
Giuseppe Minervini https://orcid.org/0000-0002-8309-1272

## References

1. Papapanou PN, Sanz M, Buduneli N, et al. Periodontitis: Consensus report of workgroup 2 of the 2017 World Workshop on the Classification of Periodontal and Peri-Implant Diseases and Conditions. *J Clin Periodontol*. 2018;45 Suppl 20:S162–S170. doi:10.1111/jcpe.12946
2. Shrivastava D, Natoli V, Srivastava KC, et al. Novel approach to dental biofilm management through guided biofilm therapy (GBT): A review. *Microorganisms*. 2021;9(9):1966. doi:10.3390/microorganisms9091966
3. Aas JA, Paster BJ, Stokes LN, Olsen I, Dewhirst FE. Defining the normal bacterial flora of the oral cavity. *J Clin Microbiol*. 2005;43(11):5721–5732. doi:10.1128/JCM.43.11.5721-5732.2005
4. Shrivastava D, Srivastava KC, Dayakara JK, et al. Bactericidal activity of crevicular polymorphonuclear neutrophils in chronic periodontitis patients and healthy subjects under the influence of areca nut extract: An in vitro study. *Appl Sci*. 2020;10(14):5008. doi:10.3390/app10145008
5. Hajishengallis G, Darveau RP, Curtis MA. The keystone-pathogen hypothesis. *Nat Rev Microbiol*. 2012;10(10):717–725. doi:10.1038/nrmicro2873
6. Bibi T, Khurshid Z, Rehman A, Imran E, Srivastava KC, Shrivastava D. Gingival crevicular fluid (GCF): A diagnostic tool for the detection of periodontal health and diseases. *Molecules*. 2021;26(5):1208. doi:10.3390/molecules26051208
7. Socransky SS, Haffajee AD, Cugini MA, Smith C, Kent RL. Microbial complexes in subgingival plaque. *J Clin Periodontol*. 1998;25(2):134–144. doi:10.1111/j.1600-051x.1998.tb02419.x
8. Darveau RP. Periodontitis: A polymicrobial disruption of host homeostasis. *Nat Rev Microbiol*. 2010;8(7):481–490. doi:10.1038/nrmicro2337
9. Nimbulkar G, Garacha V, Shetty V, et al. Microbiological and clinical evaluation of Neem gel and chlorhexidine gel on dental plaque and gingivitis in 20–30 years old adults: A randomized parallel-armed, double-blinded controlled trial. *J Pharm Bioallied Sci*. 2020;12(Suppl 1):S345–S351. doi:10.4103/jpbs.JPBS_101_20
10. Shrivastava D, Srivastava KC, Ganji KK, Alam MK, Al Zoubi I, Sghaireen MG. Quantitative assessment of gingival inflammation in patients undergoing nonsurgical periodontal therapy using photometric CIELab analysis. *Biomed Res Int*. 2021;2021:1–8. doi:10.1155/2021/6615603
11. Alam MK, Kamal MA, Srivastava KC. Special issue on current concepts and challenges in oral health: Implications for the global population. *Appl Sci*. 2023;13(5):3140. doi:10.3390/app13053140
12. Marsh PD. Dental plaque as a biofilm and a microbial community – implications for health and disease. *BMC Oral Health*. 2006;6 Suppl 1(Suppl 1):S14. doi:10.1186/1472-6831-6-S1-S14
13. Jenkinson HF, Lamont RJ. Oral microbial communities in sickness and in health. *Trends Microbiol*. 2005;13(12):589–595. doi:10.1016/j.tim.2005.09.006
14. Socransky SS, Gibbons RJ, Dale AC, Bortnick L, Rosenthal E, Macdonald JB. The microbiota of the gingival crevice area of man. I. Total microscopic and viable counts and counts of specific organisms. *Arch Oral Biol*. 1963;8:275–280. doi:10.1016/0003-9969(63)90019-0
15. Yadalam PK, Arumuganainar D, Anegundi RV, et al. CRISPR-Cas-based adaptive immunity mediates phage resistance in periodontal red complex pathogens. *Microorganisms*. 2023;11(8):2060. doi:10.3390/microorganisms11082060
16. Caselli E, Fabbri C, D'Accolti M, et al. Defining the oral microbiome by whole-genome sequencing and resistome analysis: The complexity of the healthy picture. *BMC Microbiol*. 2020;20(1):120. doi:10.1186/s12866-020-01801-y
17. Lourenço TGB, Heller D, Silva-Boghossian CM, Cotton SL, Paster BJ, Colombo APV. Microbial signature profiles of periodontally healthy and diseased patients. *J Clin Periodontol*. 2014;41(11):1027–1036. doi:10.1111/jcpe.12302
18. Jung YJ, Jun HK, Choi BK. *Porphyromonas gingivalis* suppresses invasion of *Fusobacterium nucleatum* into gingival epithelial cells. *J Oral Microbiol*. 2017;9(1):1320193. doi:10.1080/20002297.2017.1320193

19. Han YW. *Fusobacterium nucleatum*: A commensal-turned pathogen. *Curr Opin Microbiol*. 2015;23:141–147. doi:10.1016/j.mib.2014.11.013
20. Brennan CA, Garrett WS. *Fusobacterium nucleatum* – symbiont, opportunist and mycobacterium. *Nat Rev Microbiol*. 2019;17(3):156–166. doi:10.1038/s41579-018-0129-6
21. Chaushu S, Wilensky A, Gur C, et al. Direct recognition of *Fusobacterium nucleatum* by the NK cell natural cytotoxicity receptor NKp46 aggravates periodontal disease. *PLoS Pathog*. 2012;8(3):e1002601. doi:10.1371/journal.ppat.1002601
22. Bhan A, Mandal SS. Long noncoding RNAs: Emerging stars in gene regulation, epigenetics and human disease. *ChemMedChem*. 2014;9(9):1932–1956. doi:10.1002/cmdc.201300534
23. Qu S, Yang X, Li X, et al. Circular RNA: A new star of noncoding RNAs. *Cancer Lett*. 2015;365(2):141–148. doi:10.1016/j.canlet.2015.06.003
24. Hirata H, Hinoda Y, Shahryari V, et al. Long noncoding RNA MALAT1 promotes aggressive renal cell carcinoma through Ezh2 and interacts with miR-205. *Cancer Res*. 2015;75(7):1322–1331. doi:10.1158/0008-5472.CAN-14-2931
25. Liu B, Sun L, Liu Q, et al. A cytoplasmic NF-κB interacting long noncoding RNA blocks IκB phosphorylation and suppresses breast cancer metastasis. *Cancer Cell*. 2015;27(3):370–381. doi:10.1016/j.ccell.2015.02.004
26. Yuan JH, Yang F, Wang F, et al. A long noncoding RNA activated by TGF-β promotes the invasion–metastasis cascade in hepatocellular carcinoma. *Cancer Cell*. 2014;25(5):666–681. doi:10.1016/j.ccr.2014.03.010
27. Li Z, Rana TM. Decoding the noncoding: Prospective of lncRNA-mediated innate immune regulation. *RNA Biol*. 2014;11(8):979–985. doi:10.4161/rna.29937
28. Mirza AH, Kaur S, Brorsson CA, Pociot F. Effects of GWAS-associated genetic variants on lncRNAs within IBD and T1D candidate Loci. *PLoS One*. 2014;9(8):e105723. doi:10.1371/journal.pone.0105723
29. Yan B, Yao J, Liu JY, et al. lncRNA-MIAT regulates microvascular dysfunction by functioning as a competing endogenous RNA. *Circ Res*. 2015;116(7):1143–1156. doi:10.1161/CIRCRESAHA.116.305510
30. Deng T, Yang L, Zheng Z, et al. Calcitonin gene-related peptide induces IL-6 expression in RAW264.7 macrophages mediated by mmu_circRNA_007893. *Mol Med Rep*. 2017;16(6):9367–9374. doi:10.3892/mmr.2017.7779
31. Kong Z, Wan X, Zhang Y, et al. Androgen-responsive circular RNA circSMARCA5 is up-regulated and promotes cell proliferation in prostate cancer. *Biochem Biophys Res Commun*. 2017;493(3):1217–1223. doi:10.1016/j.bbrc.2017.07.162
32. Lai Z, Yang Y, Yan Y, et al. Analysis of co-expression networks for circular RNAs and mRNAs reveals that circular RNAs hsa_circ_0047905, hsa_circ_0138960 and has-circRNA7690-15 are candidate oncogenes in gastric cancer. *Cell Cycle*. 2017;16(23):2301–2311. doi:10.1080/15384101.2017.1380135
33. Cui X, Niu W, Kong L, et al. hsa_circRNA_103636: Potential novel diagnostic and therapeutic biomarker in major depressive disorder. *Biomark Med*. 2016;10(9):943–952. doi:10.2217/bmm-2016-0130
34. Zhou M, Zhao H, Wang Z, et al. Identification and validation of potential prognostic lncRNA biomarkers for predicting survival in patients with multiple myeloma. *J Exp Clin Cancer Res*. 2015;34(1):102. doi:10.1186/s13046-015-0219-5
35. Han YW, Shi W, Huang GT, et al. Interactions between periodontal bacteria and human oral epithelial cells: *Fusobacterium nucleatum* adheres to and invades epithelial cells. *Infect Immun*. 2000;68(6):3140–3146. doi:10.1128/IAI.68.6.3140-3146.2000
36. Lee HR, Rhyu IC, Kim HD, et al. In-vivo-induced antigenic determinants of *Fusobacterium nucleatum* subsp. nucleatum. *Mol Oral Microbiol*. 2011;26(2):164–172. doi:10.1111/j.2041-1014.2010.00594.x
37. Bui FQ, Johnson L, Roberts JA, et al. *Fusobacterium nucleatum* infection of gingival epithelial cells leads to NLRP3 inflammasome-dependent secretion of IL-1β and the danger signals ASC and HMGB1. *Cell Microbiol*. 2016;18(7):970–981. doi:10.1111/cmi.12560
38. Kostic AD, Chun E, Robertson L, et al. *Fusobacterium nucleatum* potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. *Cell Host Microbe*. 2013;14(2):207–215. doi:10.1016/j.chom.2013.07.007
39. Dharmani P, Strauss J, Ambrose C, Allen-Vercoe E, Chadee K. *Fusobacterium nucleatum* infection of colonic cells stimulates MUC2 mucin and tumor necrosis factor alpha. *Infect Immun*. 2011;79(7):2597–2607. doi:10.1128/IAI.05118-11
40. Hung SC, Huang PR, Almeida-da-Silva CLC, Atanasova KR, Yilmaz O, Ojcius DM. NLRX1 modulates differentially NLRP3 inflammasome activation and NF-κB signaling during *Fusobacterium nucleatum* infection. *Microbes Infect*. 2018;20(9–10):615–625. doi:10.1016/j.micinf.2017.09.014
41. Li M, Liang C. LncDC: A machine learning-based tool for long non-coding RNA detection from RNA-Seq data. *Sci Rep*. 2022;12(1):19083. doi:10.1038/s41598-022-22082-7
42. Sghaireen MG, Al-Smadi Y, Al-Qerem A, et al. Machine learning approach for metabolic syndrome diagnosis using explainable data-augmentation-based classification. *Diagnostics (Basel)*. 2022;12(12):3117. doi:10.3390/diagnostics12123117
43. Hajian-Tilaki K. Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian J Intern Med*. 2013;4(2):627–635. PMID:24009950.
44. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*. 2015;10(3):e0118432. doi:10.1371/journal.pone.0118432
45. Bolstad AI, Jensen HB, Bakken V. Taxonomy, biology, and periodontal aspects of *Fusobacterium nucleatum*. *Clin Microbiol Rev*. 1996;9(1):55–71. doi:10.1128/CMR.9.1.55
46. Patini R, Staderini E, Lajolo C, et al. Relationship between oral microbiota and periodontal disease: A systematic review. *Eur Rev Med Pharmacol Sci*. 2018;22(18):5775–5788. doi:10.26355/eurrev_201809_15903
47. Kolenbrander PE. Oral microbial communities: Biofilms, interactions, and genetic systems. *Annu Rev Microbiol*. 2000;54:413–437. doi:10.1146/annurev.micro.54.1.413
48. Xu M, Yamada M, Li M, Liu H, Chen SG, Han YW. FadA from *Fusobacterium nucleatum* utilizes both secreted and nonsecreted forms for functional oligomerization for attachment and invasion of host cells. *J Biol Chem*. 2007;282(34):25000–25009. doi:10.1074/jbc.M611567200
49. Témoin S, Wu KL, Wu V, Shoham M, Han YW. Signal peptide of FadA adhesin from *Fusobacterium nucleatum* plays a novel structural role by modulating the filament's length and width. *FEBS Lett*. 2012;586(1):1–6. doi:10.1016/j.febslet.2011.10.047
50. Kumar VS, Kumar PR, Yadalam PK, et al. Machine learning in the detection of dental cyst, tumor, and abscess lesions. *BMC Oral Health*. 2023;23(1):833. doi:10.1186/s12903-023-03571-1
51. Tan J, Li X, Zhang L, Du Z. Recent advances in machine learning methods for predicting LncRNA and disease associations. *Front Cell Infect Microbiol*. 2022;12:1071972. doi:10.3389/fcimb.2022.1071972
52. Mangan DF, Lopatin DE. Polyclonal activation of human peripheral blood B lymphocytes by *Fusobacterium nucleatum*. *Infect Immun*. 1983;40(3):1104–1111. doi:10.1128/iai.40.3.1104-1111.1983
53. Toussi DN, Liu X, Massari P. The FomA porin from *Fusobacterium nucleatum* is a Toll-like receptor 2 agonist with immune adjuvant activity. *Clin Vaccine Immunol*. 2012;19(7):1093–1101. doi:10.1128/CVI.00236-12
54. Chen S, Su T, Zhang Y, et al. *Fusobacterium nucleatum* promotes colorectal cancer metastasis by modulating KRT7-AS/KRT7. *Gut Microbes*. 2020;11(3):511–525. doi:10.1080/19490976.2019.1695494
55. Zhang S, Li C, Liu J, et al. *Fusobacterium nucleatum* promotes epithelial–mesenchymal transiton through regulation of the lncRNA MIR4435-2HG/miR-296-5p/Akt2/SNAI1 signaling pathway. *FEBS J*. 2020;287(18):4032–4047. doi:10.1111/febs.15233
56. Özveren N, Sevinç B, Sarıalioğlu Güngör A, Baltacı E, Serindere G, Özgür Ö. Evaluation of knowledge and awareness about teledentistry among dentists and patients living in Turkey. *Dent Med Probl*. 2023;60(4):593–599. doi:10.17219/dmp/150834
57. Motie P, Mohaghegh S, Kouhestani F, Motamedian SR. Effect of mechanical forces on the behavior of osteoblasts: A systematic review of in vitro studies. *Dent Med Probl*. 2023;60(4):673–686. doi:10.17219/dmp/151639

58. Golob Deeb J, Reddy N, Kitten T, Carrico CK, Grzech-Leśniak K. Viability of bacteria associated with root caries after Nd:YAG laser application in combination with various antimicrobial agents: An in vitro study. *Dent Med Probl*. 2023;60(4):649–655. doi:10.17219/dmp/171690
59. Bommala M, Koduganti RR, Panthula VR, et al. Efficacy of root coverage with the use of the conventional versus laser-assisted flap technique with platelet-rich fibrin in class I and class II gingival recession: A randomized clinical trial. *Dent Med Probl*. 2023;60(4):583–592. doi:10.17219/dmp/150887
60. Chantsalnyam T, Siraj A, Tayara H, Chong KT. ncRDense: A novel computational approach for classification of non-coding RNA family by deep learning. *Genomics*. 2021;113(5):3030–3038. doi:10.1016/j.ygeno.2021.07.004
61. Dunkel H, Wehrmann H, Jensen LR, Kuss AW, Simm S. MncR: Late integration machine learning model for classification of ncRNA classes using sequence and structural encoding. *Int J Mol Sci*. 2023;24(10):8884. doi:10.3390/ijms24108884