Journal of Big Data

RESEARCH

Open Access

Feature selection of the respiratory microbiota associated with asthma

Reem Salman[1], Ayman Alzaatreh[1*] and Mohammad T. Al Bataineh[2]

*Correspondence:
aalzaatreh@aus.edu

[1] Department of Mathematics and Statistics, American University of Sharjah, 26666 Sharjah, United Arab Emirates
[2] Center for Biotechnology, Department of Molecular Biology and Genetics, College of Medicine and Health Sciences, Khalifa University, 127788 Abu Dhabi, United Arab Emirates

Abstract

The expanding development of data mining and statistical learning techniques have enriched recent efforts to understand and identify metagenomics biomarkers in airways diseases. In contribution to the growing microbiota research in respiratory contexts, this study aims to characterize respiratory microbiota in asthmatic patients (pediatrics and adults) in comparison to healthy controls, to explore the potential of microbiota as a biomarker for asthma diagonosis and prediction. Analysis of 16 S-ribosomal RNA gene sequences reveals that respiratory microbial composition and diversity are significantly different between asthmatic and healthy subjects. Phylum *Proteobacteria* represented the predominant bacterial communities in asthmatic patients in comparison to healthy subjects. In contrast, a higher abundance of *Moraxella* and *Alloiococcus* was more prevalent in asthmatic patients compared to healthy controls. Using a machine learning approach, 57 microbial markers were identified and used to characterize notable microbiota composition differences between the groups. Among the selected OTUs, *Moraxella* and *Corynebacterium* genera were found to be more enriched on the pediatric asthmatics (p-values < 0.01). In the era of precision medicine, the discovery of the respiratory microbiota associated with asthma can lead to valuable applications for individualized asthma care.

Keywords: Asthma, Respiratory microbiota, Metagenomics, Machine learning

Background

Asthma is a widespread, long-term respiratory condition affecting more than 358 million people worldwide [26]. It is one of the most common non-communicable diseases in adults and the single most common chronic disease among children [7, 54]. Asthma is mainly caused by underlying inflammation in the lung airways, triggered by various stimuli such as viral infections, dust, smoke, animal fur, and tree pollen. Symptoms of asthma may include trouble breathing, wheezing, coughing and tightness in the chest, whereas asthma attacks are characterized by progressively increasing symptoms and severe breathing difficulties [56]. Although many patients may not show signs between attacks, untreated asthma attacks can become fatal, even in mild cases [12]. Overall, the underlying pathogenesis mechanisms of asthma remain poorly understood, though genetic studies have been increasingly able to identify more gene markers and loci related to asthma susceptibility [53]. Several risk factors involving gene-gene interactions

and gene-environment interactions are believed to influence asthma disease susceptibility [37]. In epigenetics, the identification of tissue and cell types that are best suited for the analysis remains a topic of ongoing research [34].

The word microbiom can be used to define "the ecological community of commensal, symbiotic, and pathogenic microorganisms that literally share our body space" [4]. This definition covers all microbes, including viruses, bacteria, fungi, archaea and non-fungal microscopic eukaryotes [59]. In recent decades, a large body of research has strongly suggested that the human microbiota plays a critical role in human health and disease [25, 46, 63, 73]. The gut microbiota and gut microbiota-derived metabolites, as well as their impact on host metabolism and immunology, have been the focus of most studies thus far [15, 23, 27, 48, 55, 70]. Recent research on microbial communities in other bodily locations, such as the respiratory tract, has revealed that the microbiota plays an even bigger impact in human health [41, 47, 65]. Chronic respiratory diseases such as asthma, cystic fibrosis, bronchiectasis, and chronic obstructive pulmonary disease have all been recently linked to identified changes in the microbiota composition or abundance [13, 20].

Using culture-independent 16 S rRNA sequencing, a recent study revealed variations in the microbial communities found in various regions of the healthy respiratory tract compared to those identified in the respiratory tract of asthma and chronic obstructive pulmonary disease patients (COPD) [44]. In children, similar alterations in the makeup and function of the upper airway microbiota have been linked to the exacerbation of asthma and other respiratory conditions [38, 50]. In individuals with severe asthma and similar phenotypes, certain microbiota were linked to and may influence inflammatory processes [33]. In addition, the composition and diversity of the airway microbiota were shown to be substantially associated with bronchial hyperresponsiveness in patients with subotpimally controlled asthma [32]. The relative abundance of certain phylotypes, such as those from the *Comamonadaceae*, *Sphingomonadaceae*, *Oxalobacteraceae*, and other bacterial families, was shown to be strongly associated with the degree of bronchial hyperresponsiveness [6]. Longitudinal alterations in the nasal airway microbiota were also found to mediate the impact of early antibiotic exposure on increased risk of asthma [69]. Across multiple works, *Proteobacteria* appears to be the most prevalent respiratory phylum in patients with asthma relative to non-asthmatic controls [30, 49, 74].

It is not difficult to see that asthma is a multifaceted illness with several distinct phenotypes and endotypes. Although respiratory microbiota was shown to influence on asthma development, phenotype, and severity; the exact cellular and molecular processes underlying these associations have yet to be fully elucidated [10]. As such, an important topic for current research is developing better ways to distinguish between significant asthma subtypes at the population and individual patient levels [60]. Finding relevant biomarkers and developing a better knowledge of the underlying processes associated with this disease makes it possible to advance towards improved treatment options. For this purpose, supervised machine learning algorithms can prove advantageous.

Over the past decade, the implementation of machine algorithms in the classification and analysis of microbiota and relevant biomarkers has become increasingly more

popular. Due to their compatibility with sparse datasets and high dimensional problems, supervised machine learning algorithms provide effective ways of analyzing microbiome data at efficient computational costs. Whereas the most prevalent techniques to biomarker discovery so far have been taxonomic relative abundance analysis, microbial diversity assessments using alpha and beta diversity measures, and beta pattern investigations utilizing clustering and principal coordinates analyses; these techniques can be limited in their ability to categorize unlabeled data and/or extract meaningful features from complex datasets [39]. In contrast, many well-known learning algorithms can be be used as suitable alternatives in overcoming these issues. A recent study by Wang et al. [71] employed a supervised machine learning-based approach for the classification of Autism Spectrum Disorder (ASD) based on virulence factor-related gut microbiota (VFGM) genes and Immunoglobulin A levels. Other works have similarly employed machine learning techniques for the analysis of gut and salivary microbiota in the diagnosis of diseases like colorectal cancer and liver disease [2, 9, 35]. In the study of asthma, Sharma et al. [64] discovered significant correlations between specific clinical parameters and predicted bacterial functional pathways using generalized linear models and the random forest classifier. Their findings highlighted a possible relationship between asthma-related phenotypes, endotypes, and fungal and bacterial microbiota. On the other hand, results showed no significant differences in fecal microbiota composition between adult asthmatics and non-asthmatics when extremely randomized trees-based machine learning models were used to analyze the characteristics of the intestinal microbiota [43]. In children, an unsupervised machine leaarning appraoch revealed that altered longitudinal patterns in the nasal microbiota contributed to a higher chance of developing asthma [68].

Building on previous research in which the composition, diversity, and functionality of respiratory microbiota were examined across a sample of pediatric and adult asthmatic patients [11]; this study seeks to further characterize the microbiota in asthmatic patients relative to healthy controls using a machine learning approach, while also exploring the associations between microbial markers and clinical demographics. To tackle the challenge of analyzing a sparse dataset of nearly 5853 operational taxonomic units (OTUs) for only 40 subjects, we utilized a combination of machine learning and advanced statistical techniques to identify a collection of the most relevant OTUs. By utilizing wrapper techniques, a bootstrap framework for aggregating within and between feature selection methods, and validating our findings using four popular classification algorithms; we were able to identify 57 biomarkers that characterized notable microbiota composition differences between healthy controls and asthmatic patients, including age-specific alterations. This study not only provides a noteworthy example of how big data analysis can be applied in Metagenomics but also highlights the importance of advanced tools and techniques in extracting valuable information from complex, high-dimensional data. Furthermore, our findings contribute to the growing body of knowledge about the relationship between respiratory microbiota and the diagnosis of asthma, underscoring the potential of big data analysis to shed light on complex research questions.

Methods

Participants

All participants were Emiriti residents of Sharjah, UAE. Asthma patients, defined as individuals with a current diagnosis of asthma, were recruited at the Sharjah University Hospital, UAE. Individuals who did not report having current or previous asthma, eczema or hay fever were defined as controls. Both study groups included adults and pediatrics. Demographic variables such as age, BMI, and gender were also recorded. All of the patients included for further analysis did not smoke, have any other respiratory diseases or infections, and did not use antibiotics and/or prescribed probiotics in the past 3 months. This protocol was approved by the Hospital Ethics and Research Committee, at the Sharjah University Hospital, UAE. All research participants provided their informed consent.

Sample collection and DNA extraction

Spontaneous expectorated sputum samples were obtained from all study participants. The obtained samples were spontaneous coughed up sputum (expectorated phlegm/mucous), collected after a productive cough. Sputum induction was utilized, particularly with children. All samples were collected and stored in a sterile sputum container, stored into liquid nitrogen and then transferred to -80 °C for additional analysis. Sequenced 16Sv4 amplicons were generated from the DNA samples on a MiSeq system. The mothur software package was then used to filter, identify and cluster optimal sequences into operational taxonomic units (OTUs). For the purposes of this research, ecological analysis was carried out on the sample-by-taxon abundance matrix. The raw sequences utilized for this study have been uploaded to the UCI Machine Learning Repository [3].

Traditional approach

In the first part of the analysis, traditional ecological assessment tools were employed. These methods tend to rely on investigating the distribution of microbial diversity within and between samples, in addition to analyzing the differential abundance of taxa among samples belonging to either of the investigated groups. Due to the nature of sample-by-taxon abundance matrices, zero counts can make up for a large proportion of the input matrix in these analyses [31]. As a result, the investigation of microbial communities using traditional approaches tends to be impeded by dataset sparness. To counteract the issue in this research, spurious OTUs were first filtered out of the data. Then, the investigation was carried out through a comprehensive examination of the compositional and biodiversity alpha and beta indices. Finally, differential analysis using linear discriminant analysis with effect size (LEfSe) was implemented.

In detail, alpha and beta-diversity analyses were implemented between the asthmatic patients and healthy control group. The alpha diversity values included species diversity indices (Shannon and Simpson) and species richness indices (Ace and Chao). A normalized OTU abundance table was used for the beta diversity analysis, estimated by Bray-Curtis Dissimilitudes and visualized using Principal Coordinates Analysis (PCoA). To identify dominant taxa between the groups, linear discriminant analysis with effect size (LEfSe) was used. Inside this implementation, the non-parametric factorial

Kruskal–Wallis (KW) sum-rank test is first implemented to detect features with significant differential abundance with respect to the class of interest; and biological significance is then investigated using a series of pairwise tests among subclasses using the (unpaired) Wilcoxon rank-sum test. Finally, the effect size of each differentially abundant feature is approximated by using linear discriminant analysis scores [62].

Machine learning approach

In the second part of the analysis, a machine learning approach was used to identify discriminative microbial markers across the asthmatic and healthy control groups. In ecological research, the structure of the input matrix can make supervised learning challenging to implement. Due to the high number of features in the sample-by-taxon abundance matrix, classification methods might overfit the data, resulting in models that cannot generalize well to new observations. For this reason and to lower the computational costs, reducing the number of features is often recommended. In this work, Wilcoxon-rank sum test (with FDR adjustments) was used to find top 1000 OTUs with the most significant differences between the two groups. Next, a number of feature selection algorithms were implemented to obtain a subset of the most informative OTUs out of these 1000.

Using a five-fold cross-validation precedure, a tuned Random forest model was first used to obtain the most important OTUs based on the mean decreased Gini Coefficient. In addition, an extension of Support Vector Machine Recursive Feature Elimination (namely multiple SVM-RFE), which uses resampling techniques at each iteration of a five-fold cross-validation to stabilize its feature rankings, was also utilized to obtain discriminative biomarkers. Finally, we applied a score-based ensemble feature selection framework based on bootstrap-induced diversity [61]. Within this approach, results were aggregated within and between multiple feature selection methods. Due to their computational efficiency, the following four traditional filter techniques were used inside the ensemble: Information Gain (IG), Symmertric Uncertainty (SU), Minimum Redundancy Maximum Relevance (MRMR), and Chi-Squared method (CS). To this end, 500 bootstrap samples were generated to obtain feature importance scores from each bootstrap; the importance scores were then aggregated within each single feature selection method (WAM) and between the different feature selection methods (BAM). Per its simplicity and efficiency, the aggregation was done using arithmetic mean. The ensemble was thus implemented to increase the robustness of the feature selection process and improve the accuracy of the predictions. These frameworks are illustrated in Additional file 1: Fig. S1.

Once all feature rankings were obtained, union of sets was used to combine the most important OTUs across the Random Forest, SVM-RFE, aggregated IG, aggregated CS, aggregated MRMR, aggregated CS, and the Between-Aggregation (BAM) results. Three different thresholds were used for the union: union of the 5 most important OTUs across all sets, union of the 10 most important OTUs across all sets, and union of the 20 most important OTUs across all sets. Upon comparing the unions with the individual feature selection methods, the final number of selected microbiota was based on Area Under the Receiver Operator Curve (AUROC) [28], using the four classifiers: Logistic Regression [24], Naive Bayes [36], Random Forest, and Support Vector Machine [16]. In datasets such as the one used in this work, the goal is not to predict the label classifications

of the data, but to use supervised learning to construct descriptive models that would aid in explaining the relationships between features (OTUs) in the sample-by-taxon abundance matrix. In this manner, our aim is to identify a small, but highly discriminitve subset of OTUs from the thousands of profiled genes and to utilize this subset for further study. Nonetheless, evaluation of the accuracy performance of the descriptive models will still be valuable for validating the quality of the feature selection process.

Statistical analysis

Given the small sample size, Fisher's exact test [22] was used to compare any qualitative data in the analysis. All differences in the continuous data were compared using the Mann–Whitney *U*-test [52], or the Kruskal–Wallis test [51]. The Bonferroni technique was used to adjust the p-value after multiple comparisons for the false discovery rate (FDR). All tests were two-sided, and p-value < 0.05 was considered statistically significant. All statistical analyses were performed with the relevant packages in R. LEfSe was performed using the relevant module in the Galaxy web platform [1].

Results

Clinical characteristics of the recruited subjects

In total, 40 spontaneous expectorated sputum samples were collected from 21 asthmatic patients and 19 healthy controls. Table 1 present quantitative descriptives for each of the subject groups. The mean age of asthma patients was 28 years. and 32 years for healthy subjects. The mean BMI of asthma patients was 22.10 Kg/m^2, and 26.57 Kg/m^2 for healthy subjects. Comparisons between the two groups using Mann–Whitney *U* test revealed no significant differences in age or BMI (adjusted p-values 0.5066 and 0.1759 respectively). Fisher's exact test likewise revealed no statistically significant association between the groups and gender (p-value = 1). When accounting for pediatric and adult subjects, it was not surprising to observe signifcant differences between the four groups (Adult Asthma, Adult Healthy, Pediatric Asthma, Pediatric Healthy) under age and BMI (adjusted p-values < 0.001). Pairwise Mann–Whitney comparisons adjusted using Benferroni attributed the significance to differences in age between the adults and pediatrics, as expected (adjusted p-values < 0.005), and to differences in age between the healthy and asthmatic adult subjects (adjusted p-value=0.005). Finally, the significant differences in BMI were attributed to significant differences between the adult asthmatics,

Table 1 Description of quantitative participant characteristics

	Adult asthma (N=10)		Adult healthy (N=10)		Pediatric asthma (N=11)		Pediatric healthy (N=9)	
	Mean	St. Dev	Mean	St. Dev	Mean	St. Dev	Mean	St. Dev
Age	63.90	12.66	40.00	10.55	6.73	4.15	8.00	3.12
BMI	31.47	6.71	25.26	4.73	21.58	7.25	18.24	5.23
Gender (M%, F%)	30:70		20:80		45:55		56:44	
Animal exposure (yes %)	0%		20%		9%		11%	

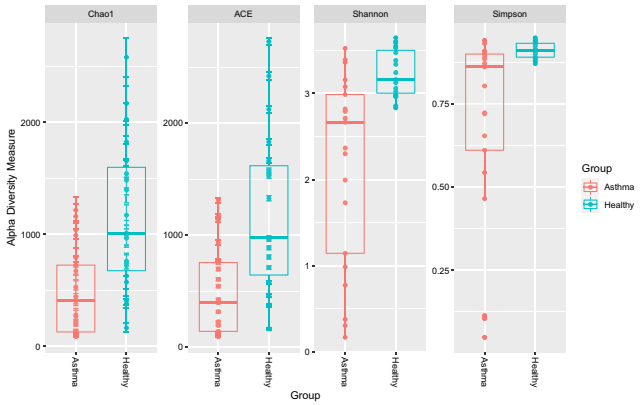
and either of the asthmatic pediatrics or healthy pediatrics (adjusted p-values 0.029 and 0.04 respectively). However, gender differences between the four groups were still not considered significant (p-value = 0.4128). For a full description of the recorded demographics, see Additional file 1: Table S1.

Sequencing characteristics and diversity

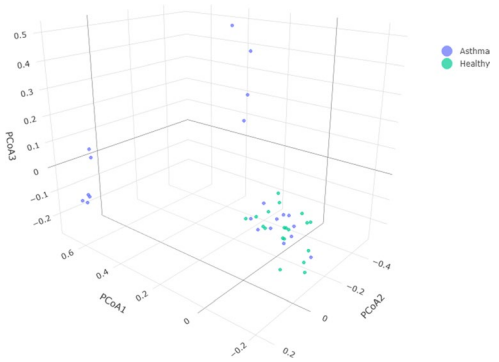
Bacterial DNA was extracted from the spontaneous expectorated sputum samples using 16 S rDNA sequencing. The 16Sv4 amplicons generated from DNA samples were sequenced on a MiSeq to obtain a dataset with 5853 OTUs. Per sample, an average of 42,657 quality-filtered reads were generated. Greengenes was used as the reference database to classify the Bacteria taxonomic composition generated from the obtained high-quality reads. Finally, OTUs were aggregated into each taxonomic rank. To address the difference in sequencing effort across samples (i.e. total numbers of sequences per sample being largely different), we applied a proportional transformation function and based our analysis on the proportional abundance of each species. The number of OTUs in each of the two groups asthmatic and healthy was 2760 and 5079, respectively. Around 24% of the 5853 OTUS overlapped between the two sample groups.

In Additional file 1: Figs. S2A, S2B, we identify the most abundant 20 OTUs across the recorded sample groups and visualize their relative abundance at the genus-level and phyla-level. Each bar length represents the mean fraction abundance of that OTU among the normalized samples in the same group. The unfilled portion of the bar represents unclassified or lower-abundance OTUs. At the Phylum-level, *Firmicutes*, *Bacteroidetes*, *Proteobacteria*, *Fusobacteria*, and *Actinobacteria* were the most abundant entities in the respiratory microbiota (in this order). Additionally, our findings agree with previous works that *Proteobacteria* seems to be the most prevalent respiratory phylum in patients with asthma in comparison to healthy controls. We also note some differences in the relative abundance of *Actinobacteria* and *Fusobacteria*. At the genus-level, asthmatic subjects showed particular abundance in *Moraxella*, *Alloiococcus*, and *Staphylococcus*; whereas healthy subjects showed greater relative abundance across the *Prevotella*, *Porphyromonas*, *Fusobacterium*, and *Veillonella* genera.

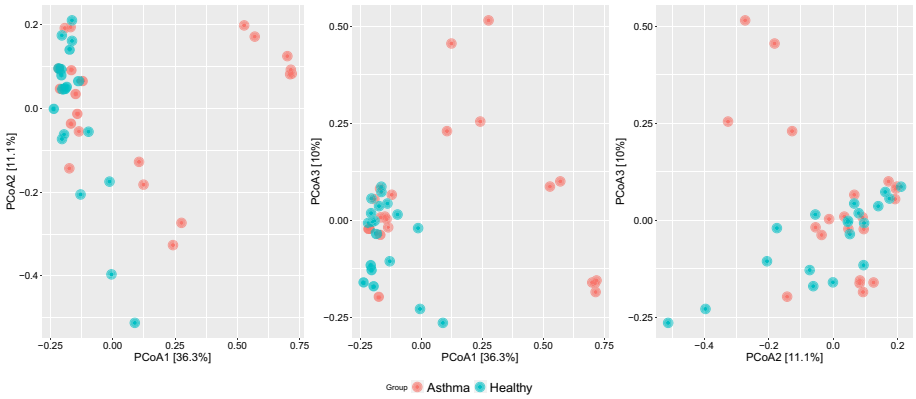
To study the variations in the respiratory microbiota across the groups, alpha diversity metrics were calculated and represented in Fig. 1A. According to the Chao and Ace diversity indices, the mean community richness is significantly lower in the asthmatic groups (adjusted p-values < 0.001) than in the healthy controls. Moreover, according to the Shannon and Simpson diversity indices, the microbial diversity was significantly lower in asthmatic subjects than in the healthy controls (adjusted p-vlaue < 0.001 and p-value = 0.002 respectively). Using Bray-Curtis dissimilarities, the normalized OTU abundances were aggregated and visualized using Principal Coordinates Analysis (PCoA) as seen in Figs. 1B, C. At the three-dimensional space (Fig. 1B), it appears larger positive scores on PCoA1 characterize a cluster of asthmatic subjects. This is further reinforced by the PCoA1 vs PCOA2 and PCoA1 vs PCoA3 plots in Fig. 1C. On the other hand, several asthmatic and healthy control subjects cluster around zero or relatively low values for PCoA2 and PCoA3. Nevertheless, Analysis of Similarities (ANOSIM) revealed



(A) Alpha diversity measures in the samples



(B) PCoA of Bray-Curtis distances across the samples in 3D-space



(C) PCoA of Bray-Curtis distances across the samples in 2D-space

Fig. 1 Overview of alpha and beta diversity analysis of respiratory microbiota between asthmatic patients and healthy subjects. **A** The graph depicts the alpha diversity boxplots for Chao1, ACE, Shannon and Simpson diversity indices distinguishing asthmatic patients and healthy control subjects. Plotted are interquartile ranges (boxes) and lines in the middle of the boxes are medians. Statistical analysis performed with paired Wilcoxon tests reveals significantly higher microbial diversity in healthy subjects (adjusted p-values< 0.05). **B**, **C** The plots depict Principal Coordinates Analysis based on the Bray-Curtis index between the above-mentioned groups in the 2D-plane and 3D-space. Some clusters of asthmatic patients can be observed on high values of PCoA1. Statistical analysis was performed with Analysis of Similarities (ANOSIM), and p-values were significant ($P = 0.007$, $R = 0.1145$)

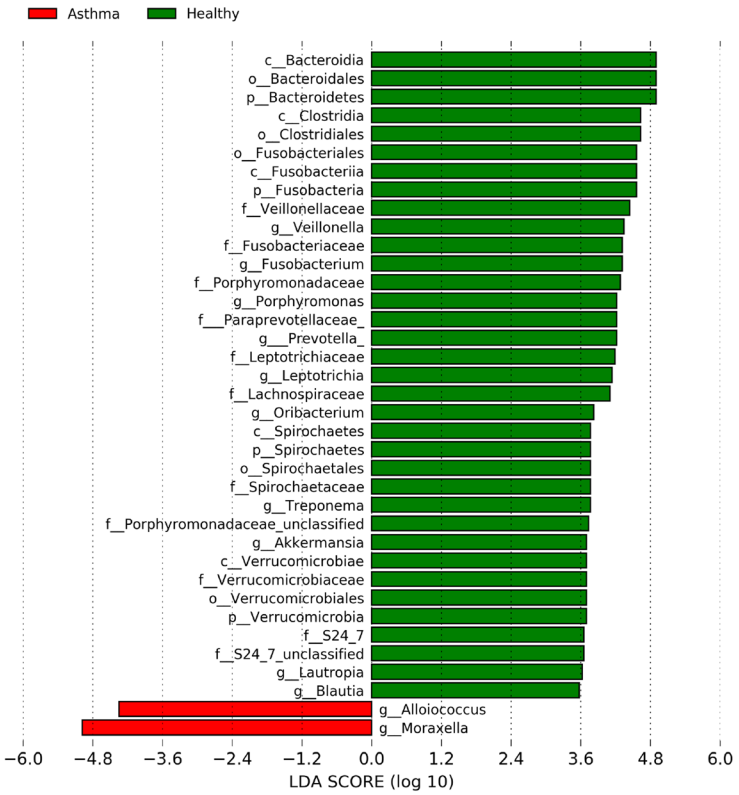
significant variations in beta-diversity across asthmatic and healthy groups ($P = 0.007$, $R = 0.1145$).

A Linear discriminant effect size analysis (LEfSe) was used to identify differences in microbial composition between asthmatic patients and healthy control groups. By combining standard tests for statistical significance with other tests concerning effect relevance and biological consistency, LEfSe determines the OTUs most likely to explain differences between classes. At the genus level, LDA scores showed significant differences in microbiota composition between asthmatic patients and healhty controls. Using a threshold of absolute LDA score 3.5 and $p < 0.02$, further analysis showed that *Moraxella* and *Alloiococcus* were more abundant in the asthmatic group, while *Veillonella*, *Fusobacterium*, *Porphyromonas*, *Prevotella*, *Leptotrichia*,*Oribacterium*, *Treponema*, *Akkermansia*, *Lautropia*, and *Blautia* genera were enriched in the healthy control group. In Fig. 2B, the bacterial tree at the center point is extended to each ring, which represents the next lower taxonomic level from phylum to genus. Yellow circles indicate non-significant differences, whereas green and red circles indicate significant differences between healthy and asthmatic subjects.

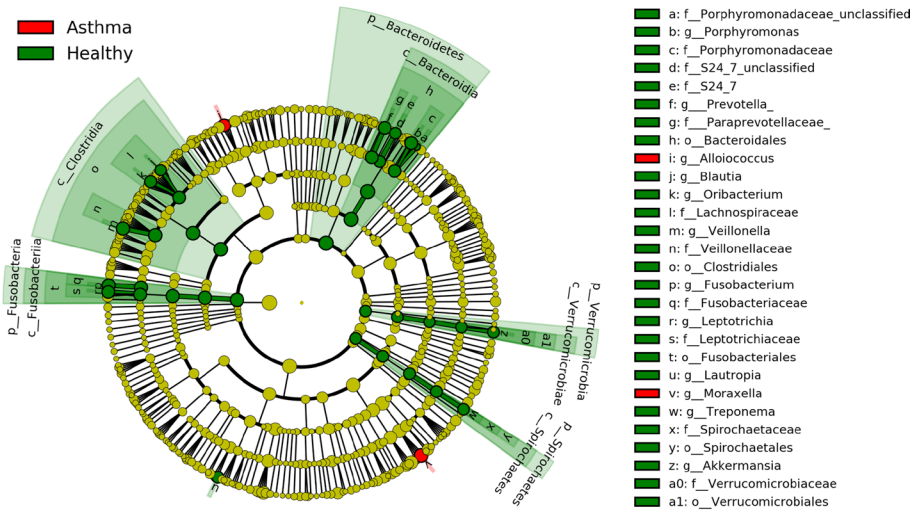
The cladogram demonstrates that many species (shown in yellow) were common between the healthy and asthmatic subjects, but there were also some distinct differences. According to the cladogram, only certain bacterial genera among *Moraxella* and *Alloiococcus* were more abundant in the asthmatic group. Meanwhile, *Bacteroidetes* (including the class *Bacteroidia*; the order *Bacteroidales*; and the families *Paraprevotellaceae*, *Porphyromonadaceae*, and *S24_7*), *Verruomicrobia* (including the class *Verrucomicrobiae*, the order *Verrucomicrobiales*, and the family *Verrucomicrobiaceae*), *Spirochaetes* (including the class *Spirochaetes*, the order *Spirochaetales*, and family *Spirochaetaceae*), *Fusobacteria* (including the class *Fusobacterila*, the order *Fusobacteriales*, and the families *Leptotrichiaceae* and *Fusobacteriaceae*), and lastly *Firmicutes* (including the class *Clostridia*, the order *Clostridiales*, and the families *Veillonellaceae* and *Lachnospiraceae*) were all more abundant in the healthy control group. Despite no significant differences between the two groups, some *Proteobacteria* genera were also more enriched in the healthy subjects.

Classification predictors

To distinguish the most discriminative respiratory microbiota in inferencing asthmatic and control groups, a machine learning feature selection approach was used. First, a Random Forest (RF) model was implemented for distinguishing between the two groups, with the most predominant genera identified on the basis of importance scores using the mean of the class-specific decreases in Gini Coefficient (Additional file 1: Fig. S3A). Next, a backward elimination procedure similar to that used in Recursive Feature Elimination for Support Vector Machines (SVM-RFE), computed the feature ranking scores at each step using a statistical analysis of the weight vectors of multiple linear SVMs within a five-fold cross-validation framework [19]. Thus, two separate OTU ranked lists were obtained. Finally, four filter feature selection methods (Information Gain, Symmetric Uncertainty, Chi-Squared test, Minimum Redundancy Maximum Relevance) were utilized within the feature selection ensemble described in the Methods section, yielding four Within-Aggregated (WAM) and one Between-Aggregated (BAM) feature selection



(A) Microbial taxa with significant differences (LDA score > 3.5)



(B) Cladogram showing the differences in relative abundance of taxa at five levels between Asthmatic (red) and Healthy (green) subjects

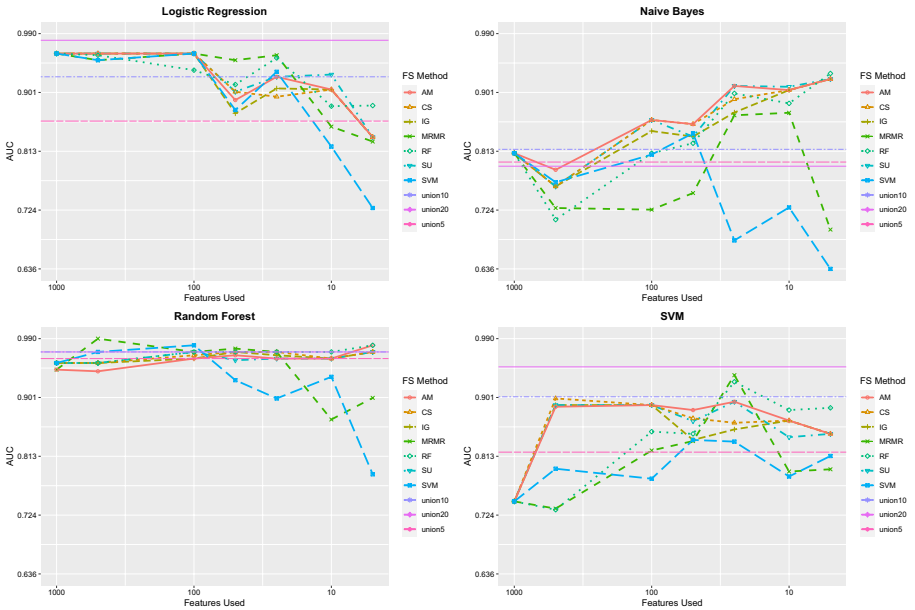
Fig. 2 Linear discriminant effect size analysis (LEfSe) between asthmatic patients and healthy control groups. LDA score> 3.5 and p = 0.05 were used in the analysis. Each identification is provided on the right (red) for ashmatic patients and on the left (green) for healthy controls, while the relevant pathways, or the names of the bacterial biomarkers, are also displayed for each entry. Most significantly differential bacterial genera was enriched on the healthy control subjects, with *Moraxella* and *Alloiococcus* being more abundant in the asthmatic group

results. These five ranked lists are hereby referred to as IG, SU, CS, MRMR, and AM for the arithmetic mean based between-aggregation. Union of the seven ranked lists (IG, SU, CS, MRMR, AM, SVM-RFE and RF) was done using different thresholds, identifying the 20, 10, and 5 most important OTUs in each of the ranked lists. Per the feature section threshold, we will refer to these sets as union20, union10, and union5.

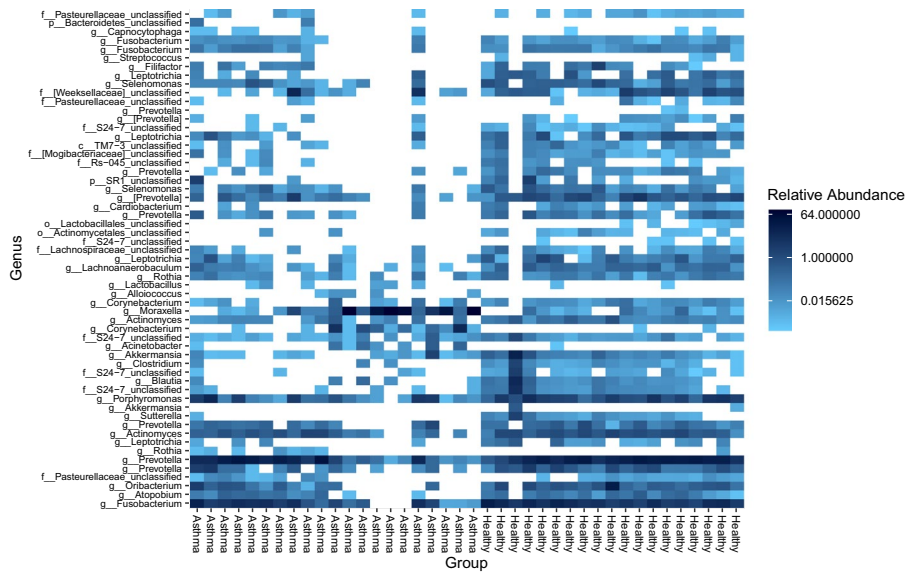
To assess the effectiveness of each of the sets in comparison to the individual ranked lists, and to determine the final selection of the most important OTUs; we tested all selected feature sets against the four classifiers: Logistic Regression, Naive Bayes, Random Forest, and Support Vector Machine (SVM). The model classification procedure was done within a five-fold cross-validation loop. Then, the accuracy of the predictive models was determined using AUC (Area under the Receiver Operator Curve). The obtained results are represented by Fig. 3A. Note that each of the union sets has a fixed number of features: 15 OTUs in union5, 28 OTUs in union10, and 57 OTUs in union20. This is why the union sets' accuracies are represented by horizental lines in Fig. 3A. Regarding model tuning, the Naive Bayes model used default parameters, with no Laplacian correction, and the prior probabilities of class membership were set based on the class proportions from the training set. Normal density estimation was also applied. Logistic Regression used a weakly informative default prior distribution with coefficients set to prior mean and scale of 0 and 2.5, and intercept set to prior mean and scale of 0 and 10, respectively. Degrees of freedom were set to 12 for coefficients and 1 for intercept. Random Forest utilized 500 trees with $\sqrt{p}$ randomly sampled as candidates at each split, where p denotes the total number of features. SVM utilized a radial basis function kernel with gamma set to $\frac{1}{p}$ and the cost parameter to 1. The insensitive-loss function used a 0.1 epsilon, and class weights were defaulted to 1. Based on their published performance in machine learning applications with high dimensionality, we expect these classifiers to provide a good performance benchmark. It should be highlighted, nevertheless, that for this portion of the analysis, our primary objective is not to achieve high prediction accuracy, but to use a limited set of informative OTUs to discriminate between healthy and asthmatic samples for further investigation.

In three of the classifiers, the model AUC improves upon selecting the most relevant OTUs. In line with previous work, the Random Forest classifier appears to be the strongest predictor [39]. In the Naive Bayes classifier, the arithmetic mean based between-aggregation (AM) seems to be the best identifier of OTU selections. Meanwhile in Random Forest, there is greater overlap between the accuracies derived from the feature selection results. However, in both logistic regression and SVM, it appears that the best AUC is obtained by using the union20 set. Accordingly, we believe that a union of the 20 topmost ranking features in each of the lists might be most productive. All identified OTUs ($p = 57$) are displayed at the family, genus, and species levels in Additional file 1: Table S2. Note that many of the identified OTUs in the machine learning approach agree with our findings using traditional means. To further expand on these observations, we analyze our selected OTUs and their discriminative powers with respect to their microbiota compositions.

In Fig. 3B, the displayed heatmap shows the relative abundance of each of the selected genus-level microbiota across the 40 samples. From the figure, it is clear that most of our 57 identified OTUs are more enriched in the healthy group samples



(A) Classification performance of the most important OTUs



(B) Heatmap of the selected OTUs in union50

Fig. 3 OTU feature selection results. **A** The classification performance of the most important bacterial microbiota (OTUs) selected by each of the feature selection approaches used in the study: AM, CS, IG, MRMR, and SU are based on ensemble feature selection; RF and SVM are wrapper-based feature selection techniques; Union5, Union10, and Union20 combine the feature selection results. Under four classifiers, the selected OTUs can distinguish asthmatic subjects with nearly 99% AUC accuracy. **B** A subset of the most important 57 bacteriaal microbiota (OTUs) for asthmatic diagnosis developed by the feature selection methdology (Union20). The heat map shows the relative abundance of the 57 most important bacterial microbiota (OTUs) for asthmatic diagnosis

in comparison to the asthmatic group, emphasizing the validity of our results. Mann–Whitney U-test (adjusted for FDR) reveals that the observed abundance differences are significant in 43 of the selected OTUs (see Table S2 for p-values). In particular, we note multiple significant identifiers at the genus-level such as

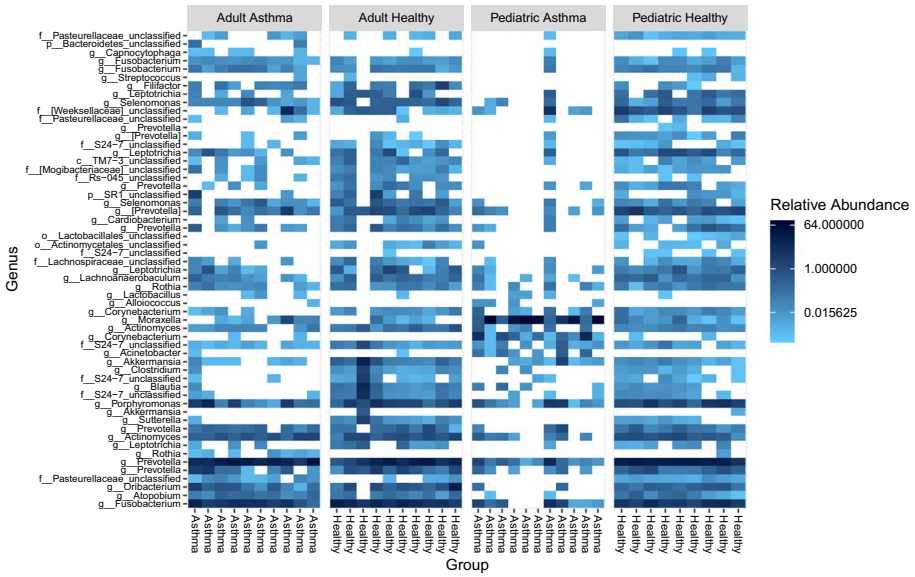
Lachnoanaerobaculum, *Sutterella*, *Oribacterium*, *Actinomyces*, *Selenomonas*, *Rothia*, *Cardiobacterium*, *Corynebacterium*, *Clostridium*, and several unclassified bacteria (adjusted p-values<0.01). In constrast, a significantly greater abundance of *Moraxella* and *Alloiococcus* (adjusted p-values 0.008 and 0.036) was prevalant among the asthmatic samples compared to the healthy subjects, consistently with previous LEfsE and microbiota abundance analysis results. At the genus-level, biomarkers which showed greater relative abundances across the asthmatic and healthy groups, such as *Moraxella*, *Alloiococcus*, *Prevotella*, *Porphyromonas*, and *Fusobacterium*, were all selected by the machine learning approach. Likewise, biomarkers identified using LEfsE, whether differentially abundant among healthy subjects such as *Leptotrichia*, *Prevotella*, *Blautia*, *Porphyromonas*, and *Akkermansia*; or among the asthmatic patients, such as *Moraxella* and *Alloiococcus*, were also selected.

Effect of age

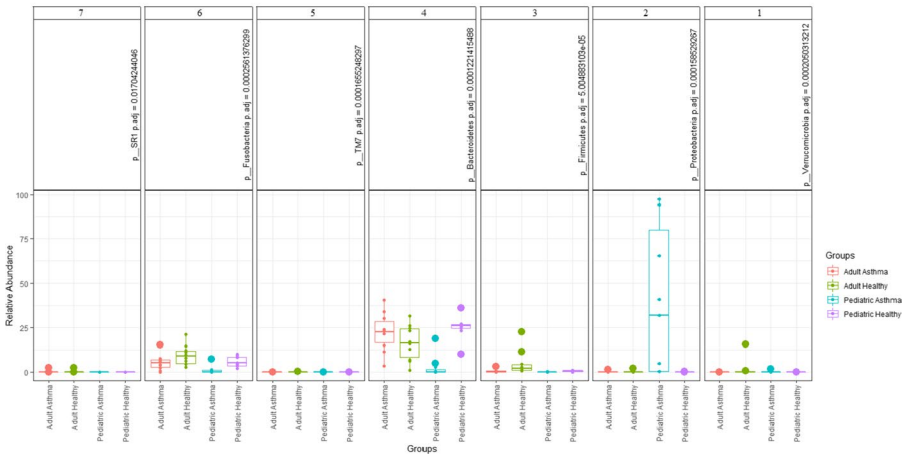
At the age level, the number of OTUs in each of the four groups Adult Asthma, Adult Healthy, Pediatric Asthma, and Pediatric Healthy was 1815, 1554, 945, and 3525 respectively. Only around 5% of the 5853 OTUS overlapped between the four sample groups. In terms of relative microbial abundance, Figure S4 reveals visible differences between the most genus-level abundant OTUS. At large, *Moraxella* is distinctly more abundant in pediatric asthmatic subjects than in any other group, whereas *Veillonella* exhibits the least abundance on pediatric asthmatics. Across adult subjects, differences between *Prevotella*, *Streptococcus*, and *Neisseria* are notable between healthy and asthmatic adults.

According to the Chao and Ace diversity indices, the mean community richness is significantly lower in the pediatric asthmatic subjects than in any other group (adjusted p-values < 0.001), and similarly less diverse on Shannon and Simpson diversity indices (adjusted p-value < 0.001 and p-value $= 0.002$ respectively). Using Bray-Curtis dissimilarities, Principal Coordinates Analysis (PCoA) in Figure S5B reveals a cluster of pediatric asthamtic subjects with higher loadings on PCoA1 and a cluster of healthy adult subjects negatively loaded on both PCoA2 and PCoA3. Analysis of Similarities (ANOSIM) also shows significant differences in beta-diversity across the different age subgroups ($P = 0.001$, $R = 0.3369$).

Based on the selected 57 OTUs using the machine learning approach, Fig. 4A reveals that *Atopobium*, *Actinomyces*, *Oribacterium*, *Prevotella*, *Fusobacterium*, and *Selenomonas* were all more prevalent across the healthy pediatrics and both healthy and asthmatic adults, but not in pediatric asthmatics (pairwise adjusted p-values<0.05). On the other hand, some bacteria such as *Moraxella* and *Corynebacterium* could be identified in significantly greater abundances among the pediatric asthmatics (pairwise adjusted p-values<0.01). In Fig. 4B, Kruskal Wallis was used to test significantly differentially abundant Phylya in the selected OTUS across the age subgroups. This included *SR1*, *Fusobacteria*, *TM7*, *Bacteroidetes*, *Firmicutes*, *Proteobacteria*, and *Verrucomicrobia*. In particular, *Proteobacteria* shows significant abundance among the pediatric asthmatic subjects. This is in line with previous results across prior subsections.



(A) Heatmap of the selected OTUs in union50 (Age Subgroup)



(B) Significantly differentially abundant Phyla in Union50 under different Age Subgroups using Kruskal-Wallis test

Fig. 4 OTU feature selection results by Age Subrgoup. **A** The heat map depicts the relative abundance of the 57 most important bacterial microbiota (OTUs) for asthma diagnosis identified using the machine learning approach. The colors depict each OTU's relative genera abundance among the tested subjects, grouped by age and ashtma diagnosis. **B** The figure depicts the Kruskual Wallis analysis of significant differential abundance of the 57 most important bacterial microbiota (OTUs) identified using the machine learning approach. The boxplots depict the relative phyla abundance of the selected OTUs among the tested subjects, grouped by age and ashtma diagnosis. Significant differential abundance can be observed among these phyla, as indicated by the adjusted p-values

Discussion

Recent technological advances and the rapid development of bioinofrmatics analysis methods have increasingly identified associations between disease and microbiota within the human body. Differences in the microbial populations in the respiratory airways have always been related to the diagnosis of asthma across many studies. In the current work, we characterize the respiratory microbiota in asthmatic versus healthy patients, with further emphasis on the subject age group (i.e. adults and pediatrics). The

results demonstrate that the respiratory microbiota's taxonomic composition and diversity were all significantly different between healthy and asthmatic samples. *Proteobacteria* especially dominated respiratory microbiota compositions in asthmatic patients at the phyla level and *Moraxella* at the genus level. Moreover, we found that multiple microbiota genera were more abundant across the healthy controls than the asthmatic patients.

As highlighted by the traditional ecological assessment methods, greater diversity and richness of respiratory microbiota could be observed in healthy samples compared with asthmatic patients. These results were supported by previous literature findings using alpha and beta diversity measures [5, 18, 29]. In line with this, it has been observed that antibiotic use in early life was associated with the development of childhood asthma, since it could lower microbial diversity and affect microbial composition [42, 72]. Several works have also shown that people who grow up in surroundings with high microbiological diversity have a much lower chance of developing asthma [14, 21, 40]. In this manner, diversified microbial environments can promote diversified human microbiota, leading to reduced risk of atopic illness development and improved lung function, especially among pediatrics [17, 41]. Using the traditional ecological approach, our findings in this work emphasize the difference in microbial composition between healthy and asthmatic patients, and further validate the results obtained by the machine learning framework. In terms of the most prevalent genus-level taxa, *Prevotella*, *Porphyromonas*, *Rothia*, were more abundant in healthy individuals, while more significant amounts of *Moraxella*, *Alloiococcus*, *Streptobacillus* were abundant across asthmatic subjects. In terms of age, pediatric asthmatics were particularly numerous in the *Moraxella* and *Alloiococcus* genera, whereas adult asthmatics had a higher enrichments of bacterial microbiota from the *Streptococcus* and *Prevotella* genera than the other groups. In general, differences between the adult subjects could be characterized across *Prevotella*, *Streptococcus*, and *Neisseria* genera, in line with previous work [41].

We characterize the most pertinent microbiota associated with discriminating asthmatic samples from healthy controls using a machine learning approach. A union of multiple ranked lists across several feature selection methods was identified. The selected subset of 57 bacterial families resulted in prediction accuracy nearly equal or even higher than that of classifiers trained on most OTUs. Upon further analysis, these OTUs provide primary evidence that the respiratory ecology in asthmatic patients differs from that of healthy people. This is supported by the findings for both the traditional approach and machine learning framework. Across both results, increased abundances of the identified *Moraxella* and *Alloiococcus* biomarkers could be observed in asthmatic patients, whereas *Prevotella*, *Porphyromonas*, *Fusobacterium*, *Leptotrichia*, *Blautia*, and *Akkermansia* genera were identified for the healthy subjects. Additional identifiers at the genus-level were uniquely considered through the machine learning approach, including *Streptococcus*, *Cardiobacterium*, *Corynebacterium*, *Clostridium*, *Sutterella*, and *Actinomyces*. According to an Australian birth cohort study, early *Streptococcus* colonization was significantly associated with a younger age of first respiratory illness and a persistent wheeze in preschool age, especially among those with early allergic sensitization. [67]. Similarly, *Prevotella* has been observed to be the most abundant genera in the lungs of healthy subjects, whereas *Staphylococcus* and *Haemophilus* were more abundant in

asthmatic patients within a sample of 47 subjects [30]. In other literature, healthy subjects showed higher abundances of *Streptococcus*, *Veillonella*, *Prevotella*, and *Neisseria* of phylum *Firmicutes* compared to asthmatic and COPD patients [57, 77], supporting the current findings.

In analyzing the effect of age, our findings noted significant *Proteobacteria* abundance among the pediatric asthmatic subjects, and significant differential abundances in other Phyla such as *Bacteroidetes* and *Fusobacteria*. Among the selected OTUs, *Moraxella* and *Corynebacterium* genera were significantly more enriched on the pediatric asthmatics. These results were in line with several studies which have found an increase in the phylum *Proteobacteria*, particularly the species *Haemophilus*, among asthmatic patients. [30, 49, 66, 74]. A recent work by Hauptmann and Schaible [29] also revealed that *Proteobacteria* was found in greater abundance in asthmatic children's airway microbiota than in healthy controls; *Bacterioidetes* was more prevalent in asthmatics overall. Our findings are also consistent with earlier studies showing reduced enrichment on *Bacteroidetes* and *Fusobacteria* in both non-severe and severe asthmatic groups in comparison to the healthy group, and *Firmicutes* showing higher enrichment in severe asthmatics [8, 74]. Among school-aged children with asthma, a longitudinal study recently revealed that a shift to *Moraxella* colonization at the Yellow Zone (where the patient's symptoms are at danger of progressing to a severe exacerbation), as well as a decreased *Corynebacterium* abundance, were both linked to an increased risk of severe exacerbations the following year [75].

On the other hand, it should be noted that the feature selection subsets which comprised of a smaller number of relevant bacterial families (i.e., 10, 5) did not perform as well under most classifiers, suggesting that too small of a species count might be inadequate for defining the respiratory microbiota associated with asthma. This discovery implies that respiratory dysbiosis in asthma is caused by a complicated interaction of many bacterial and fungal groups. Previous works has suggested that the *Moraxellaceae* family and its genus *Moraxella*, alongside three key fungal species, exhibit substantial interactions with the airway microbiota [45]. In this work and others, the ecological differences between asthmatic and non-asthmatic patients, especially children, have been largely associated with differences in *Moraxella* gene expression [18, 49, 58, 75]. These bacterial colonizations are further observed to alter the likelihood and severity of viral infections. For example, when the respiratory syncytial virus is present, an airway microbiome dominated by *Moraxella* predisposes to lower respiratory tract infections and raises the risk of fever [67]. However, the influence of environmental diversity should still be considered, as indicated by previous findings in which the link between *Moraxella* and asthma was only evident in non-farm children [18]. Longitudinal research is needed to further understand the relationship between asthma development and *Moraxella* colonization.

In conclusion, we discovered several respiratory bacterial species linked with asthmatic patients, shedding further insight on the respiratory microbiota effects on asthma pathogenesis with respect to age. Using a machine learning approach, our study identified 57 relevant microbial markers in diagnosing and characterizing asthma. Many of the findings agreed with the traditional ecological assessment methods used in this work and could further identify additional biomarkers supported by previous literature. We

show how various current supervised machine learning approaches may be used to reliably classify asthma diagnosis and select highly discriminative subsets of taxa for further exploration. This approach is beneficial when the the number of independent features or the intricacy of their interconnections make univariate hypothesis testing ineffective.

The findings emphasized here could contribute to a better knowledge of the identification and composition of the respiratory microbiota in asthmatic patients, which might affect the use of the microbiome as a treatment strategy for chronic respiratory illnesses like asthma. However, a limitation of the present study that could prevent extrapolation of the results was the small sample size, as only 40 respiratory microbiota spontaneous expectorated sputum samples were used. Due to the high dimensionality in the sample-by-taxon abundance matrix, some filtering was processed on the data prior to implementing the machine learning approach. Several strategies for reducing the amount of OTU features utilizing correlation and taxonomy information have been recently developed [76]. Performing dimensionality reduction by decreasing the phylogenetic specificity of taxonomic groupings, or leveraging the inherently hierarchical structure of the OTUs using an algorithm like hierarchical feature engineering (HFE) are other possible alternatives [39]. Alpha and beta diversity analyses of the data are also likely to provide useful features for classification so a combination of both the traditional and machine learning methods may be implemented. Finally, this study's utilization of spontaneous expectorated sputum samples in this study made it difficult to distinguish the lower respiratory tract microbiota from upper respiratory tract microbiota. Future research may traverse further deep into the relationship between respiratory microbiota and its location in the respiratory tract, in line with the microbiota community's influence on respiratory diseases like asthma.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40537-023-00767-8.

Additional file 1: Figure S1. Overview of the ensemble feature selection framework. The dataset is first divided into a training dataset, and a testing dataset. Then, multiple training subsamples are generated by bootstrapping the training dataset. Next, a number of feature selection techniques FS1, . . . , FSt are applied on each subsample, generating a feature importance score $\ell j \in R$ for every feature. The aggregation is thus two-fold; Within Aggregation Methodis used for aggregating the importance scores within a single feature selection method and Between Aggregation Methodis used for aggregating the importance scores between different feature selection methods. Once the feature set is sorted from the most to the least important, a rank vector is obtained and used to characterize the most important OTUs. **Table S1.** Description of data characteristics. **Figure S2.** Microbial abundance by Asthma Group. Phylum composition was compared among the asthmatic patients and healthy controls using the the 20 most abundant OTUs. In the graph, the relative abundance is expressed as the mean value for each group. At the Phylum level, Proteobacteria was more enriched in subjects with asthma relative to controls.At the Genus level, Moraxella, Alloiococcus, and Staphylococcus most frequently dominate the respiratory samples from asthmatic patients, whereas Prevotella, Porphyromonas, Fusobacterium, and Veillonella dominate the respiratory samples from healthy controls. **Figure S3.** Diagnostic models based on microbiota selected by a Random Forest model with 250 trees and 50 variables randomly sampled as candidates at each split. The size of terminal nodes was fixed to a minimum of 1.The most important bacterial microbiotaare listed in descending order of relevance by mean decrease in Gini Coefficient. The mean decrease in Gini coefficient is a measure of how each OTU contributes to the homogeneity of the random forest's nodes and leaves. The greater the mean decrease in Gini coefficient, the more important the feature is for the Random Forest classifier.ROC curves for the Random Forest classifier on which the OTU importance scores were obtained. The model was used to distinguish between asthmatic patients and healthy controls based on the sample-by-taxon abundance matrix. The true positive fraction is the proportion of real positivesthat are correctly classified as positive; the false positive fraction is the proportion of false positivesthat are incorrectly classified as positive. The ROC curve summarizes the true positive and false positive rates and the AUC reflects the classifier's ability to correctly differentiate between two classes. Table S2: Final selected OTUs. P-values characterize differences in the relative abundances between asthmatic and healthy subjects. **Figure S4.** Microbial abundance by Asthma and Age Subgroup. Genus composition was compared among the asthmatic pediatrics, asthmatic adults, healthy pediatrics and healthy adults using the 20 most abundant OTUs. The relative abundance is expressed as the mean

value for each group. Moraxella is distinctly more abundant in pediatric asthmatics than in any other group, whereas Veillonella, Prevotella, and Neisseria were less enriched on asthmatic pediatrics. **Figure S5.** Overview of alpha and beta diversity analysis of the respiratory microbiota in asthmatic patients and healthy controls by Age Subgroup.The graph depicts the alpha diversity boxplots for Chao1, ACE, Shannon and Simpson diversity indices distinguishing asthmatic patients and healthy control subjects by Age Subgroup. Plotted are interquartile rangesand lines in the middle of the boxes are medians. Statistical analysis performed with paired Wilcoxon tests reveals significantly lower microbial diversity in asthmatic pediatrics.-The plots depict Principal Coordinates Analysis based on the Bray-Curtis index between the above-mentioned groups in 2D-space. Some clusters of asthmatic pediatrics can be observed on high values of PCoA1. Moreover, adult healthy subjects were generally observed to have negative loadings on each PCoA. Statistical analysis was performed with Analysis of Similarities, and p-values were significant.

Acknowledgements
The authors are grateful for the comments and suggestions by the referees and the Editor. Their comments and suggestions have greatly improved the paper. The authors are also gratefully acknowledge that the work in this paper was supported, in part, by the Open Access Program from the American University of Sharjah.

Author contributions
All authors have contributed to all sections including the methodology, the data analysis, and the conclusion. All authors read and approved the final manuscript.

Funding
This work was supported by the American University of Sharjah Open Access Fund.

Availability of data and materials
The datasets generated and analysed during the current study are available in the the UCI Machine Learning Repository [3] at https://archive-beta.ics.uci.edu/dataset/795/16srna+asthma.

Declarations

Ethics approval and consent to participate
The study protocol was approved by the Hospital Ethics and Research Committee at the Sharjah University Hospital, UAE. Informed consent was obtained from all individual participants included in the study.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 26 June 2022 Accepted: 17 May 2023
Published online: 01 June 2023

References

1. Afgan E, Baker D, Batut B, Van Den Beek M, Bouvier D, Čech M, Chilton J, Clements D, Coraor N, Grüning BA, et al. The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic acids research. 2018;46(W1):537–44.
2. Ai L, Tian H, Chen Z, Chen H, Xu J, Fang J-Y. Systematic evaluation of supervised classifiers for fecal microbiota-based prediction of colorectal cancer. Oncotarget. 2017;8(6):9546.
3. Al Bataineh M. 16sRNA. UCI Mach Learn Repos. 2023. https://doi.org/10.1186/s12879-020-05427-3.
4. Al Bataineh MT, Henschel A, Mousa M, Daou M, Waasia F, Kannout H, Khalili M, Kayasseh MA, Alkhajeh A, Uddin M, et al. Gut microbiota interplay with covid-19 reveals links to host lipid metabolism among middle eastern populations. Front Microbiol. 2021;12:3197.
5. Al Bataineh MT, Künstner A, Dash NR, Abdulsalam RM, Al-Kayyali RZA, Adi MB, Alsafar HS, Busch H, Ibrahim SM. Corrigendum: altered composition of the oral microbiota in depression among cigarette smokers-a pilot study. Front Psychiatry. 2023;14:1175044. https://doi.org/10.3389/fpsyt.2023.1175044.
6. Alharbi KS, Alenezi SK, Alnasser SM. Microbiome in asthma. In: Gupta G, Oliver BG, Dua K, Singh A, MacLoughlin R, editors. Microbiome in inflammatory lung diseases. Singapore: Springer; 2022. p. 65–77. https://doi.org/10.1007/978-981-16-8957-4_5.
7. Asher I, Pearce N. Global burden of asthma among children. Int J Tuberc Lung Dis. 2014;18(11):1269–78.
8. Ballarini S, Rossi GA, Principi N, Esposito S. Dysbiosis in pediatrics is associated with respiratory infections: is there a place for bacterial-derived products? Microorganisms. 2021;9(2):448. https://doi.org/10.3390/microorganisms9020448.
9. Bang S, Yoo D, Kim S-J, Jhang S, Cho S, Kim H. Establishment and evaluation of prediction model for multiple disease classification based on gut microbial data. Sci Rep. 2019;9(1):1–9.
10. Barcik W, Boutin RC, Sokolowska M, Finlay BB. The role of lung and gut microbiota in the pathology of asthma. Immunity. 2020;52(2):241–55.

11. Bataineh MTA, Hamoudi RA, Dash NR, Ramakrishnan RK, Almasalmeh MA, Sharif HA, Al-Hajjaj MS, Hamid Q. Altered respiratory microbiota composition and functionality associated with asthma early in life. BMC Infect Dis. 2020;20(1):1–11.
12. Bork K, Anderson JT, Caballero T, Craig T, Johnston DT, Li HH, Longhurst HJ, Radojicic C, Riedl MA. Assessment and management of disease burden and quality of life in patients with hereditary angioedema: a consensus report. Allergy Asthma Clin Immunol. 2021;17:1–14.
13. Budden KF, Shukla SD, Rehman SF, Bowerman KL, Keely S, Hugenholtz P, Armstrong-James DP, Adcock IM, Chotirmall SH, Chung KF, et al. Functional effects of the microbiota in chronic respiratory disease. Lancet Respir Med. 2019;7(10):907–20.
14. Campbell B, Lodge C, Lowe A, Burgess J, Matheson M, Dharmage S. Exposure to 'farming' and objective markers of atopy: a systematic review and meta-analysis. Clin Exp Allergy. 2015;45(4):744–57.
15. Carding S, Verbeke K, Vipond DT, Corfe BM, Owen LJ. Dysbiosis of the gut microbiota in disease. Microb Ecol Health Dis. 2015;26(1):26191.
16. Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol. 2011;2:27–12727.
17. Coscia A, Bardanzellu F, Caboni E, Fanos V, Peroni DG. When a neonate is born, so is a microbiota. Life. 2021;11(2):148. https://doi.org/10.3390/life11020148.
18. Depner M, Ege MJ, Cox MJ, Dwyer S, Walker AW, Birzele LT, Genuneit J, Horak E, Braun-Fahrländer C, Danielewicz H, et al. Bacterial microbiota of the upper respiratory tract and childhood asthma. J Allergy Clin Immunol. 2017;139(3):826–34.
19. Duan K-B, Rajapakse JC, Wang H, Azuaje F. Multiple svm-rfe for gene selection in cancer classification with expression data. IEEE Trans Nanobioscience. 2005;4(3):228–34.
20. Dumas A, Bernard L, Poquet Y, Lugo-Villarino G, Neyrolles O. The role of the lung microbiota and the gut-lung axis in respiratory infectious diseases. Cell Microbiol. 2018;20(12):12966.
21. Ege MJ, Mayer M, Normand A-C, Genuneit J, Cookson WO, Braun-Fahrländer C, Heederik D, Piarroux R, von Mutius E. Exposure to environmental microorganisms and childhood asthma. N Engl J Med. 2011;364(8):701–9.
22. Fisher RA. Statistical methods for research workers. In: Kotz S, Johnson NL, editors. Breakthroughs in statistics. New York: Springer; 1992. p. 66–70.
23. Flint HJ, Scott KP, Louis P, Duncan SH. The role of the gut microbiota in nutrition and health. Nat Rev Gastroenterol Hepatol. 2012;9(10):577–89.
24. Gelman A, Jakulin A, Pittau MG, Su Y-S. A weakly informative default prior distribution for logistic and other regression models. Ann Appl Stat. 2008;2(4):1360–83.
25. Gill SR, Pop M, DeBoy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE. Metagenomic analysis of the human distal gut microbiome. Science. 2006;312(5778):1355–9.
26. Gruffydd-Jones K, Thomas M, Roman-Rodríguez M, Infantino A, FitzGerald JM, Pavord I, Haddon JM, Elsasser U, Vogelberg C. Asthma impacts on workplace productivity in employed patients who are symptomatic despite background therapy: a multinational survey. J Asthma Allergy. 2019;12:183–94.
27. Gu S, Chen Y, Wu Z, Chen Y, Gao H, Lv L, Guo F, Zhang X, Luo R, Huang C, et al. Alterations of the gut microbiota in patients with coronavirus disease 2019 or h1n1 influenza. Clin Infect Dis. 2020;71(10):2669–78.
28. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29–36.
29. Hauptmann M, Schaible UE. Linking microbiota and respiratory disease. FEBS Lett. 2016;590(21):3721–38.
30. Hilty M, Burke C, Pedro H, Cardenas P, Bush A, Bossley C, Davies J, Ervine A, Poulter L, Pachter L, et al. Disordered microbial communities in asthmatic airways. PLoS ONE. 2010;5(1):8578.
31. Hu T, Gallins P, Zhou Y-H. A zero-inflated beta-binomial model for microbiome data analysis. Stat. 2018;7(1):185.
32. Huang YJ, Nelson CE, Brodie EL, DeSantis TZ, Baek MS, Liu J, Woyke T, Allgaier M, Bristow J, Wiener-Kronish JP, et al. Airway microbiota and bronchial hyperresponsiveness in patients with suboptimally controlled asthma. J Allergy Clin Immunol. 2011;127(2):372–81.
33. Huang YJ, Nariya S, Harris JM, Lynch SV, Choy DF, Arron JR, Boushey H. The airway microbiome in patients with severe asthma: associations with disease features and severity. J Allergy Clin Immunol. 2015;136(4):874–84.
34. Hudon Thibeault A-A, Laprise C. Cell-specific DNA methylation signatures in asthma. Genes. 2019;10(11):932.
35. Iwasawa K, Suda W, Tsunoda T, Oikawa-Kawamoto M, Umetsu S, Takayasu L, Inui A, Fujisawa T, Morita H, Sogo T, et al. Dysbiosis of the salivary microbiota in pediatric-onset primary sclerosing cholangitis and its potential as a biomarker. Sci Rep. 2018;8(1):1–10.
36. John GH, Langley P. Estimating continuous distributions in bayesian classifiers. arXiv preprint. 2013. arXiv:1302.4964.
37. Kabesch M, Tost J. Recent findings in the genetics and epigenetics of asthma and allergy. Semin Immunopathol. 2020;42:43–60.
38. Kim B-S, Lee E, Lee M-J, Kang M-J, Yoon J, Cho H-J, Park J, Won S, Lee S, Hong S. Different functional genes of upper airway microbiome associated with natural course of childhood asthma. Allergy. 2018;73(3):644–52.
39. Knights D, Costello EK, Knight R. Supervised classification of human microbiota. FEMS Microbiol Rev. 2011;35(2):343–59.
40. ...Kopp MV, Muche-Borowski C, Abou-Dakn M, Ahrens B, Beyer K, Blümchen K, Bubel P, Chaker A, Cremer M, Ensenauer R, Gerstlauer M, Gieler U, Hübner I-M, Horak F, Klimek L, Koletzko BV, Koletzko S, Lau S, Lob-Corzilius T, Nemat K, Peters EMJ, Pizzulli A, Reese I, Rolinck-Werninghaus C, Rouw E, Schaub B, Schmidt S, Steiß J-O, Striegel AK, Szépfalusi Z, Schlembach D, Spindler T, Taube C, Trendelenburg V, Treudler R, Umpfenbach U, Vogelberg C, Wagenmann M, Weißenborn A, Werfel T, Worm M, Sitter H, Hamelmann E. S3 guideline allergy prevention*. Allergol Select. 2022;6:61–97. https://doi.org/10.5414/ALX02303E.
41. Koppen IJ, Bosch AA, Sanders EA, van Houten MA, Bogaert D. The respiratory microbiota during health and disease: a paediatric perspective. Pneumonia. 2015;6(1):90–100.

42. Kozyrskyj AL, Ernst P, Becker AB. Increased risk of childhood asthma from antibiotic use in early life. Chest. 2007;131(6):1753–9.
43. Kullberg RF, Haak BW, Abdel-Aziz MI, Davids M, Hugenholtz F, Nieuwdorp M, Galenkamp H, Prins M, Maitland-van der Zee AH, Wiersinga WJ. Gut microbiota of adults with asthma is broadly similar to non-asthmatics in a large population with varied ethnic origins. Gut Microb. 2021;13(1):1995279.
44. Lee HW, Sim YS, Jung JY, Seo H, Park J-W, Min KH, Lee JH, Kim B-K, Lee MG, Oh Y-M, et al. A multicenter study to identify the respiratory pathogens associated with exacerbation of chronic obstructive pulmonary disease in Korea. Tuberculosis Respir Dis. 2022;85(1):37.
45. Liu H-Y, Li C-X, Liang Z-Y, Zhang S-Y, Yang W-Y, Ye Y-M, Lin Y-X, Chen R-C, Zhou H-W, Su J. The interactions of airway bacterial and fungal communities in clinically stable asthma. Front Microbiol. 2020;11:1647.
46. Liu Q, Duan ZP, Ha DK, Bengmark S, Kurtovic J, Riordan SM. Synbiotic modulation of gut flora: effect on minimal hepatic encephalopathy in patients with cirrhosis. Hepatology. 2004;39(5):1441–9.
47. Man WH, de Steenhuijsen Piters WA, Bogaert D. The microbiota of the respiratory tract: gatekeeper to respiratory health. Nat Rev Microbiol. 2017;15(5):259–70.
48. Manichanh C, Borruel N, Casellas F, Guarner F. The gut microbiota in IBD. Nat Rev Gastroenterol Hepatol. 2012;9(10):599–608.
49. Marri PR, Stern DA, Wright AL, Billheimer D, Martinez FD. Asthma-associated differences in microbial composition of induced sputum. J Allergy Clin Immunol. 2013;131(2):346–52.
50. McCauley K, Durack J, Valladares R, Fadrosh DW, Lin DL, Calatroni A, LeBeau PK, Tran HT, Fujimura KE, LaMere B, et al. Distinct nasal airway bacterial microbiotas differentially relate to exacerbation in pediatric patients with asthma. J Allergy Clin Immunol. 2019;144(5):1187–97.
51. McKight PE, Najab J. Kruskal–Wallis test. Cors Encycl Psychol. 2010. pp. 1–1.
52. McKnight PE, Najab J. Mann-whitney U test. Cors Encycl Psychol. 2010. pp. 1–1.
53. Ntontsi P, Photiades A, Zervas E, Xanthou G, Samitas K. Genetics and epigenetics in asthma. Int J Mol Sci. 2021;22(5):2412.
54. Organization WHO et al. Assessing national capacity for the prevention and control of noncommunicable diseases: report of the 2019 global survey. 2020.
55. O'Toole PW, Jeffery IB. Gut microbiota and aging. Science. 2015;350(6265):1214–5.
56. Papi A, Brightling C, Pedersen SE, Reddel HK. Asthma. Lancet. 2018;391(10122):783–800. https://doi.org/10.1016/S0140-6736(17)33311-1.
57. Park H, Shin JW, Park S-G, Kim W. Microbial communities in the upper respiratory tract of patients with asthma and chronic obstructive pulmonary disease. PLoS ONE. 2014;9(10): 109710.
58. Pérez-Losada M, Castro-Nallar E, Bendall ML, Freishtat RJ, Crandall KA. Dual transcriptomic profiling of host and microbiota during health and disease in pediatric asthma. PLoS ONE. 2015;10(6):0131819.
59. Rogers GB, Shaw D, Marsh RL, Carroll MP, Serisier DJ, Bruce KD. Respiratory microbiota: addressing clinical questions, informing clinical practice. Thorax. 2015;70(1):74–81.
60. Saglani S, Custovic A. Childhood asthma: advances using machine learning and mechanistic studies. Am J Respir Crit Care Med. 2019;199(4):414–22.
61. Salman R, Alzaatreh A, Sulieman H, Faisal S. A bootstrap framework for aggregating within and between feature selection methods. Entropy. 2021;23(2):200.
62. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, Huttenhower C. Metagenomic biomarker discovery and explanation. Genome Biol. 2011;12(6):1–18.
63. Sekirov I, Russell SL, Antunes LCM, Finlay BB. Gut microbiota in health and disease. Physiol Rev. 2010. https://doi.org/10.1152/physrev.00045.2009.
64. Sharma A, Laxman B, Naureckas ET, Hogarth DK, Sperling AI, Solway J, Ober C, Gilbert JA, White SR. Associations between fungal and bacterial microbiota of airways and asthma endotypes. J Allergy Clin Immunol. 2019;144(5):1214–27.
65. Soltani S, Zakeri A, Zandi M, Kesheh MM, Tabibzadeh A, Dastranj M, Faramarzi S, Didehdar M, Hafezi H, Hosseini P, et al. The role of bacterial and fungal human respiratory microbiota in covid-19 patients. BioMed Res Int. 2021. https://doi.org/10.1155/2021/667079.
66. Tarquinio KM, Karsies T, Shein SL, Beardsley A, Khemani R, Schwarz A, Smith L, Flori H, Karam O, Cao Q, Haider Z, Smirnova E, Serrano MG, Buck GA, Willson DF. Airway microbiome dynamics and relationship to ventilator-associated infection in intubated pediatric patients. Pediatr Pulmonol. 2022;57(2):508–18. https://doi.org/10.1002/ppul.25769.
67. Teo SM, Mok D, Pham K, Kusel M, Serralha M, Troy N, Holt BJ, Hales BJ, Walker ML, Hollams E, et al. The infant nasopharyngeal microbiome impacts severity of lower respiratory infection and risk of asthma development. Cell Host Microb. 2015;17(5):704–15.
68. Toivonen L, Karppinen S, Schuez-Havupalo L, Waris M, He Q, Hoffman KL, Petrosino JF, Dumas O, Camargo CA, Hasegawa K, et al. Longitudinal changes in early nasal microbiota and the risk of childhood asthma. Pediatrics. 2020;146(4): e20200421.
69. Toivonen L, Schuez-Havupalo L, Karppinen S, Waris M, Hoffman KL, Camargo CA Jr, Hasegawa K, Peltola V. Antibiotic treatments during infancy, changes in nasal microbiota, and asthma development: population-based cohort study. Clin Infect Dis. 2021;72(9):1546–54.
70. Valdes AM, Walter J, Segal E, Spector TD. Role of the gut microbiota in nutrition and health. BMJ. 2018;361: k2179.
71. Wang M, Doenyas C, Wan J, Zeng S, Cai C, Zhou J, Liu Y, Yin Z, Zhou W. Virulence factor-related gut microbiota genes and immunoglobulin a levels as novel markers for machine learning-based classification of autism spectrum disorder. Comput Struct Biotechnol J. 2021;19:545–54.
72. Wang Z, Bafadhel M, Haldar K, Spivak A, Mayhew D, Miller BE, Tal-Singer R, Johnston SL, Ramsheh MY, Barer MR, et al. Lung microbiome dynamics in COPD exacerbations. Eur Respir J. 2016;47(4):1082–92.

73. Wen L, Ley RE, Volchkov PY, Stranges PB, Avanesyan L, Stonebraker AC, Hu C, Wong FS, Szot GL, Bluestone JA, et al. Innate immunity and intestinal microbiota in the development of type 1 diabetes. Nature. 2008;455(7216):1109–13.
74. Zhang Q, Cox M, Liang Z, Brinkmann F, Cardenas PA, Duff R, Bhavsar P, Cookson W, Moffatt M, Chung KF. Airway microbiota in severe asthma and relationship to asthma severity and phenotypes. PLoS ONE. 2016;11(4):0152724.
75. Zhou Y, Jackson D, Bacharier LB, Mauger D, Boushey H, Castro M, Durack J, Huang Y, Lemanske RF, Storch GA, et al. The upper-airway microbiota and loss of asthma control among asthmatic children. Nat Commun. 2019;10(1):1–10.
76. Zhou Y-H, Gallins P. A review and tutorial of machine learning methods for microbiome host trait prediction. Front Genetics. 2019;10:579.
77. Zhu T, Jin J, Chen M, Chen Y. The impact of infection with COVID-19 on the respiratory microbiome: a narrative review. Virulence. 2022;13(1):1076–87.

Publisher's Note