

RESEARCH ARTICLE

Open Access



Classification and prediction of *Mycobacterium Avium* subsp. *Paratuberculosis* (MAP) shedding severity in cattle based on young stock heifer faecal microbiota composition using random forest algorithms

Alexander Umanets^{1,3}, Annemieke Dinkla¹, Stephanie Vastenhouw¹, Lars Ravesloot¹ and Ad P. Koets^{1,2*}

Abstract

Background: Bovine paratuberculosis is a devastating infectious disease caused by *Mycobacterium avium* subsp. *paratuberculosis* (MAP). The development of the paratuberculosis in cattle can take up to a few years and vastly differs between individuals in severity of the clinical symptoms and shedding of the pathogen. Timely identification of high shedding animals is essential for paratuberculosis control and minimization of economic losses. Widely used methods for detection and quantification of MAP, such as culturing and PCR based techniques rely on direct presence of the pathogen in a sample and have little to no predictive value concerning the disease development. In the current study, we investigated the possibility of predicting MAP shedding severity in cattle based on the faecal microbiota composition. Twenty calves were experimentally infected with MAP and faecal samples were collected biweekly up to four years of age. All collected samples were subjected to culturing on selective media to obtain data about shedding severity. Faecal microbiota was profiled in a subset of samples (n = 264). Using faecal microbiota composition and shedding intensity data a random forest classifier was built for prediction of the shedding status of the individual animals.

Results: The results indicate that machine learning approaches applied to microbial composition can be used to classify cows into groups by severity of MAP shedding. The classification accuracy correlates with the age of the animals and use of samples from older individuals resulted in a higher classification precision. The classification model based on samples from the first 12 months of life showed an AUC between 0.78 and 0.79 (95% CI), while the model based on samples from animals older than 24 months showed an AUC between 0.91 and 0.92 (95% CI). Prediction for samples from animals between 12 and 24 month of age showed intermediate accuracy [AUC between 0.86 and 0.87 (95% CI)]. In addition, the results indicate that a limited number of microbial taxa were important for classification and could be considered as biomarkers.

*Correspondence: ad.koets@wur.nl

¹ Department of Bacteriology, Host Pathogen Interaction and Diagnostics Development, Wageningen Bioveterinary Research, Lelystad, The Netherlands

Full list of author information is available at the end of the article



© The Author(s) 2021. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Conclusions: The study provides evidence for the link between microbiota composition and severity of MAP infection and shedding, as well as lays ground for the development of predictive diagnostic tools based on the faecal microbiota composition.

Keywords: *Mycobacterium avium* subsp. *paratuberculosis*, Gut microbiota, Machine learning, Prediction, Pathogen shedding, Bovine, Random forest

Background

Bovine paratuberculosis (Johne's disease) is a widespread chronic wasting disease caused by *Mycobacterium avium* subsp. *paratuberculosis* (MAP) infection. The disease is extremely relevant in the context of dairy farm practice due to induction of significant economic losses related to the animal wasting, infertility and decrease in the milk production [1].

It is generally accepted that calves can be infected with MAP in the first days of life, however, the development of MAP shedding and clinical symptoms can take years and varies substantially between individuals in severity [2]. Faecal shedding of MAP into the environment has been shown to be high in animals with clinical symptoms [3]. Nevertheless, severe MAP shedding can occur even in asymptomatic animals, which in turn is correlated with an increased rate of transmission of paratuberculosis in a herd [4]. Therefore, timely identification of high shedding animals is essential for pathogen control and minimization of economic losses.

Widely used methods for detection and quantification of MAP, such as culturing and PCR based techniques, rely on the presence of MAP cells or its DNA in a sample and have little to no predictive value for disease development. In addition, despite a very high specificity of culturing and PCR based tests, their sensitivity can vary greatly due to the intermediate shedding of MAP cells by infected animals [5, 6]. In this light, the development of a diagnostic tool for populations with an endemic MAP infection that will allow prediction of MAP shedding severity without relying on the presence of the pathogen in a sample or a strong and specific immunological response is an attractive goal.

The pathophysiology of bovine paratuberculosis is complex as exposure of animals to MAP may lead to clearance of MAP (no infection) or infection. Infection of the distal part of the small intestine as the preferred site of infection leads to the formation of granulomatous lesions characterized by accumulation of macrophages in the lamina propria. These lesions can contain large amounts of MAP cells (multibacillary) or very few (paucibacillary). In some animals these lesions remain relatively stable for years while in others a progressive disease develops which will lead to intractable diarrhoea, proteins losing enteropathy and ultimately death. The

factors driving these changes and different outcomes are only partly understood and involve multiple factors including host immunity and genetics, MAP genetic factors as well as environmental factors (reviewed in Koets et al. [7]). In cattle, as a part of the infectious process, MAP colonizes the intestine and induces changes in the intestinal environment due to invasion into the mucosa [8], and is forced to interact with members of the intestinal microbiota. It has been shown in observational studies that MAP infection induces noticeable differences in the gut microbiota composition of affected cows [9, 10]. In an experimental rabbit model alteration of microbiota by diet in turn also influenced the severity of MAP infection [11]. These findings indicate a bilateral relationship between MAP and other intestinal microorganisms.

The composition of the intestinal microbiota has been shown to be a biomarker for gastro-intestinal health and disease development. Microbiota composition analysis showed a potential for prediction of gastro-intestinal cancer [12], liver disease [13], and other not-communicative disorders [14–16]. However, the forecasting of a disease or disorder development using microbiota composition is not a trivial task. Microbiota composition in both humans and cattle have been shown to be relatively stable within a single individual [17, 18], however the composition shifts dramatically with the change of hosts diet [19] and age [20, 21] in cattle, lambs and humans. These changes introduce a high level of noise in the data that could render forecasting very difficult or inaccurate. Another challenge comes from the complex nature of microbiota composition data. Large datasets can be composed from hundreds of samples and each sample may contain a very diverse microbial community. Complex, not normally distributed, zero inflated data can be challenging for statistical analysis and deciphering of the relevant signal. Nevertheless, machine learning approaches have shown a great promise in addressing challenges presented by the microbiota compositional data including prediction of host susceptibility to pathogens [22]. Machine learning is a branch of artificial intelligence that advanced considerably in recent years and is now commonly used in multiple aspects of everyday life as well as science. Machine learning is a broad term for a family of statistical techniques that are taking advantages of a large amount of data and computational power of modern

computers to detect patterns in data [23]. Machine learning algorithms, and in particular Random Forest, became a popular solution for analysis of various types of microbiota data, due to its flexibility [24].

In the current study the feasibility of MAP shedding intensity prediction based on microbial composition differences between animals with different cumulative shedding of MAP cells throughout their life was investigated. The following study is a proof of concept study that aimed to lay the foundation for development of microbiota composition based diagnostic and prediction tools.

Methods

Animals

Details of the animal experiment have been previously described in detail in a PhD thesis [25] and in a publication by Ganusov et al. [26]. In short, twenty neonatal female Holstein–Friesian calves were purchased from several commercial high health farms in the Netherlands with a documented paratuberculosis unsuspected status and housed at the Wageningen Bioveterinary Research (WBVR) pathogen free facilities (Lelystad, The Netherlands). Calves were raised following best practices, on a diet with restricted access to a commercially available milk replacer and calf concentrate. In addition, calves had free access to hay and drinking water. Following weaning at 6 weeks of age the animals were fed with a conventional diet based on grass silage, concentrates and free access to water corresponding to their age and lactation status. The diet was never supplemented with fresh grass and animals were kept indoors for the duration of the experiment.

Practices of the animal husbandry during the experiment were closely resembling dairy cattle farming practices commonly accepted in the Netherlands except for outdoor grazing. Animals were bred around 15 months of age. In the second year of the study six animals were culled due to infertility and two due to development of pathological conditions not related to the experimental MAP infection. The remaining 12 animals survived for at least 50 months out of the total of 55 months of the experiment duration. During the experiment none of the animals developed any signs of clinical paratuberculosis (severe diarrhoea, weight loss, emaciation, oedema).

Experimental infection and samples collection

A faecal suspension from a single donor-cow with clinical signs of paratuberculosis and consistently shedding MAP as confirmed by culture and MAP specific IS900 qPCR [27] was used as the inoculum. Every calf was orally dosed with 20 g of inoculum in 200 mL of commercially available milk replacer recommended for use in dairy calf rearing, three times per week for the period of four

weeks. After the experimental infection of animals, faecal samples were collected every two weeks.

Samples were collected from the rectum using a clean set of disposable latex gloves per calf without lubricant. Samples were transported to lab immediately where a part was taken for direct MAP culture and quantification, the rest was rapidly stored for further analysis at -20°C as has been described previously [25].

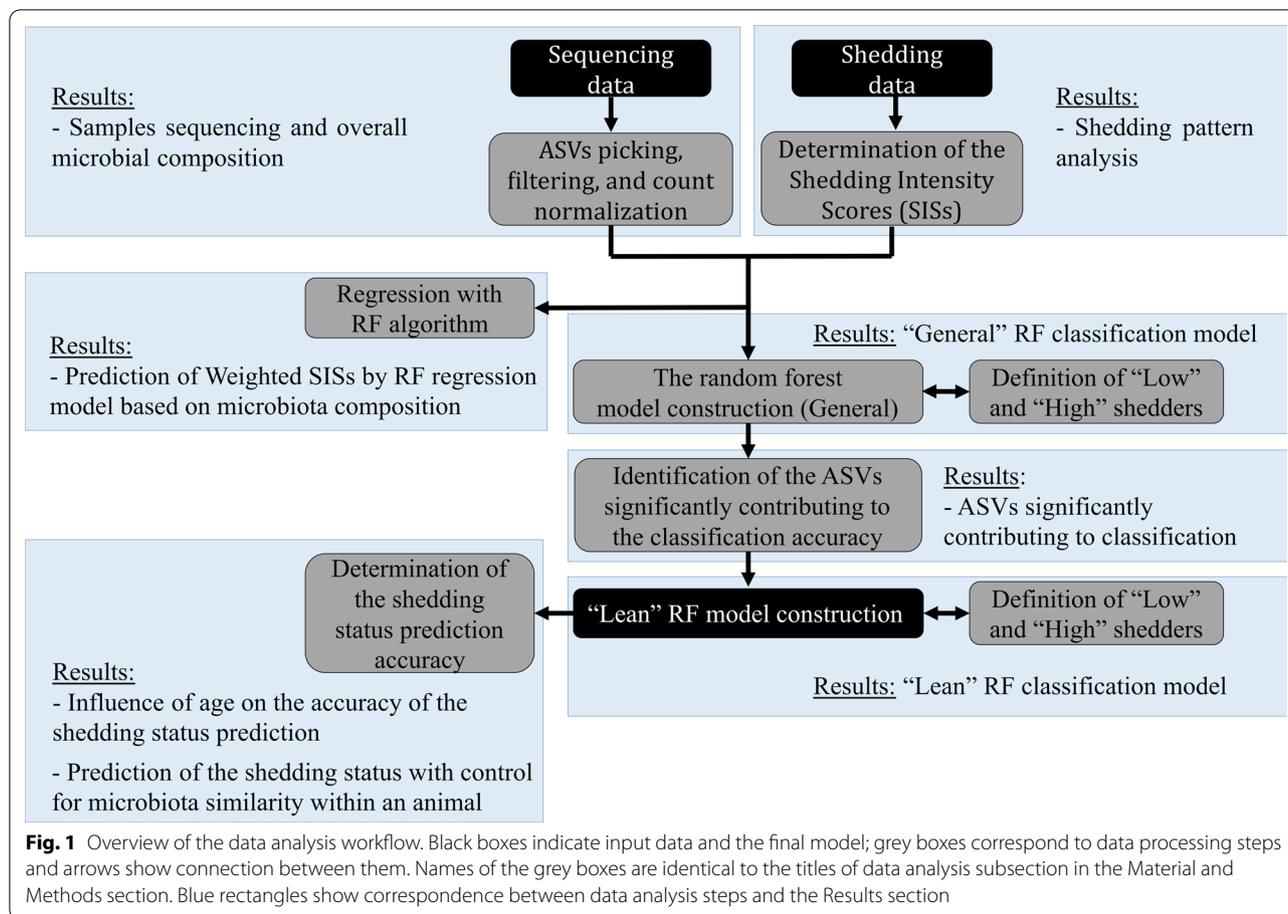
DNA extraction and amplicons library preparation

For microbiota profiling, a subset of samples was selected from the experimentally infected cattle. Two hundred sixty two samples were selected to represent the overall dataset with the particular focus on samples that were collected at crucial animal husbandry time-points: pre-weaning period (1 month), post-weaning period (3 and 7 months), start of sexual maturity (12 months) and start of first lactation (24 months); Additional file 1). Prior to DNA extraction faecal samples were thawed and 0.1 g of a sample was subjected to repetitive bead-beating (3×30 s with 5 s cooldown in between) using Lysing Matrix B (2 mL tube) and the FastPrep-24 instrument (MP Biomedicals). Total DNA was extracted individually from every faecal sample using QIAamp Fast DNA Stool Mini Kit (QIAGEN) according to the manufacture instructions. In short, following homogenization the samples were lysed in a lysis buffer, treated with InhibiTEX to remove inhibitors from the sample, and proteins were digested using a proteinase K treatment. DNA was subsequently bound to a silica spin column, washed twice and eluted in a low-salt buffer. Extracted DNA was subjected to quantification using CLARIOstar Plus (BMG Labtech) and Quant-iT™ dsDNA Assay Kit, high sensitivity (ThermoFisher Scientific) according to instructions provided by the manufacturer.

Amplicon library preparation and sequencing were performed at BaseClear (Leiden, The Netherlands). The V3-V4 region of 16 s rRNA gene was amplified with 341F (5'-CCTACGGGNGGCWGCAG-3') and 785R (5'-GAC TACHVGGGTATCTAATCC-3') [28] primers in a two-step protocol. Generated amplicons were appended with Illumina adaptors and sequenced on the Illumina MiSeq platform. The sequencing data was demultiplexed, Illumina adaptors were removed and the data was transferred from BaseClear to WBVR for further analysis.

Data analysis and statistics

A schematic outline of the data analysis steps is shown in Fig. 1. The outline shows the major steps of analysis, corresponding results and connections between them. Sub-headings of the Data analysis subsections are identical to the names of the grey coloured action boxes in Fig. 1.



The complete data analysis pipeline is available as a GitHub project at https://github.com/AlexanderUm/WBVR_MAP_Microbiota. Additional file 2 contains the samples’ metadata used for the data processing.

Determination of the shedding intensity scores

Isolation and quantification of the MAP colonies were performed according to methods published previously [25, 26]. Based on the number of colonies per tube a Shedding Intensity Score (SIS) was assigned to each sample. The SISs were assigned as following: 0—no colonies observed; 1—from 1 to 10 colonies observed; 2—from 11 to 100 colonies observed; 3—more than 100 colonies observed (Additional file 3). The SIS values reflect intensity of MAP shedding at a given time point and were used for further calculations of shedding scores. Based on the individual SIS a Weighted SIS was calculated. For calculation of Weighted SISs the sum of all SISs per animal was divided by the number of samples from that animal. The Weighted SISs were taken as an objective indication of MAP shedding severity. Weighted SISs were calculated for all available samples (I) and for samples from

the first 20 months of the experiment (II). The Weighted SISs II were calculated with the goal to mitigate effects of differences in the time an animal was present in the experiment (Additional file 4: Fig. S1A). In total 20 animals were enrolled in the experiment, however, some were culled before the end of the experiment. In total, eight animals were culled between 20 and 26 months of age (“Early Culled”) due to the inability to breed or health related problems. Weighted SISs I and Weighted SISs II in the subset of animals which survived until the end of the experiment showed a high and significant correlation (Pearson; $r=0.89$, $p<0.0001$), and further analysis were performed using Weighted SISs II (Additional file 4: Fig. S1B).

ASVs picking, filtering, and count normalization

The raw data was processed following the DADA2 v1.14.1 pipeline [29] in R v3.6.2 statistical and programming environment [30]. The final output of DADA2 pipeline is an abundance table of amplicon sequence variants (ASVs). An ASV is a unique sequence of a marker gene that approximates a microbial species or strain and could

be used as the lowest taxonomic units (for more details see publication of Callahan et al. [26]). In short, primers and low quality bases were truncated from the reads, the model of error-rate within sequences was inferred from the data, ASVs were picked and chimeric sequences removed. AVSs were taxonomy assigned as implemented in DADA2 (function “assignTaxonomy”) using the suggested version of the manually curated Silva v138 database [31]. Taxonomic and abundance tables produced by the DADA2 pipeline were combined with metadata into a phyloseq object [32]. Prior to further analysis ASVs were filtered to retain only taxa that were present in more than three samples. Also, samples with less than 1000 reads were removed. Sufficiency of the sequencing depth was assessed by rarefaction curves with a 200 bp step as implemented in the package *vegan* (v 2.5–6). Prior to further analysis, ASV count data was normalized using cumulative sum scale (CSS) transformation [33].

The random forest model construction (General)

The random forest (RF) algorithm [34] as implemented in the *randomForest* v4.6–14 package [35] was used for sample classification and regression. The RF model was fine-tuned by determining the optimal number of taxa at each split (“mtry” parameter) and the number of classification trees (“ntree” parameter). The “mtry” parameter was tuned using the *tuneRF* function (*randomForest* package) with step factor two. The optimal number of classification trees was determined by constructing the RF model five times with a different number of trees, and subsequent identification of the models with the lowest out of bag error (OOBE).

Definition of “Low” and “High” shedders

The Weighted SISs were used as a guide to assign animals as a “Low” or “High” shedder, however no obvious split between animals was observed. Therefore, animals were classified into “Low” or “High” shedders based on accuracy of RF classification model. Several RF models with various configurations of “High” and “Low” shedding groups were constructed. Samples were assigned as “High” shedders if their Weighted SIS II was above the selected cut-off value and as “Low” shedders if below. Cut-off values with a range between between 0.45 and 1.19 were used to capture shifts in the shedding intensity and a RF model was built for each available Weighted SIS II within this range.

Regression with RF algorithm

The regression model was built as implemented in the *randomForest* package. The sufficient number of trees for the RF model was identified as follows. A RF model was built using default parameters (*randomForest*) and

an excessively large number of trees (20,000). The built model was used for plotting relationships between accuracy and the number of used trees (*plot(RF)*). A total of 7501 trees were used as the final *ntree* parameter for the model with the default *mtry* parameter.

To test the prediction accuracy beyond OOBE a subtraction approach was used as described below. Consequentially several RF regression models were built, each time using a data-set with ten uniquely subtracted samples until no possibility for unique subtraction was left. RF regression models were used to predict values of the corresponding subtracted values (generic R function “predict”). To account for the random variations in RF models the model building for each set of subtracted samples was repeated five times.

Identification of the taxa significantly contributing to the classification accuracy and “Lean” RF model construction

Taxonomic features with a significant contribution to the RF accuracy were identified by a permutation test as implemented in *rfPermute* v2.1.81 package [36]. Only features with the significance levels below threshold values ($\alpha=0.05$) for the mean decrease accuracy and the mean decrease purity (Gini purity index [35]) were considered as important for classification. Significant ASVs were used as input features for building a new simplified RF model (“Lean”). The “mtry” and “ntree” parameters for the simplified RF model were optimised as described above.

Determination of the shedding status prediction accuracy

Accuracy of the shedding status prediction was assessed by construction of the receiver operating characteristic (ROC) curve and calculation of the area under the curve (AUC). To calculate ROC curve and AUC a dataset was split into “training” and “validation” datasets. The “training” dataset was used to construct the RF classification model (package *randomForest*) and shedding status of samples from the corresponding “validation” dataset was predicted (generic R function “predict”). The resulting prediction was used to construct ROC curves and calculated AUC (package *ROCR*).

To understand the influence of cows’ age on the accuracy of classification the ROC curve and AUC for different age groups were calculated. Three age groups were defined: Early ($n=82$)—before 12 months of age, Middle ($n=90$)—12 to 24 months of age and Late ($n=85$)—older than 24 months of age. The ROC curve and AUC for a specific age group was calculated based on the group specific “validation” and “training” datasets. Fifty percent of the samples from the target age group was randomly drawn to create a “validation” dataset. The samples from

the other age groups and not drawn samples from the target group comprised a “training” dataset. The overall ROC curve and AUC were estimated using 20% of randomly drawn data points from all samples as a “validation” dataset and the rest of the samples as a “training” dataset. Splitting of the target data into “validation” and “training” with consequent evaluation of the ROC curve and AUC was repeated 99 times for each age group and the complete dataset.

Next, we ruled out the possibility of samples classification into a shedding group due to similarity of microbiota within one animal during its lifespan. We used all samples from a single animal as the “validation” dataset and the rest of the samples as the “training” dataset. This procedure was repeated for every animal and the results of prediction were compared with the shedding status.

Data wrangling and visualization

For data wrangling and visualization, the package tidyverse v1.2.1 [37] was used. For visualization of heatmaps the package ComplexHeatmaps [38] was used.

Results

Shedding pattern analysis

No apparent division in “Low” and “High” shedders among animals based on the Weighted SISs was observed. A gradual increase in Weighted SISs from 0.3 (lowest observed) to 0.84, slight jump from 0.84 to 1.19 and further gradual increase to 1.41 (highest observed) was observed. The mean shedding value among all samples was 0.53 (Fig. 2A, B).

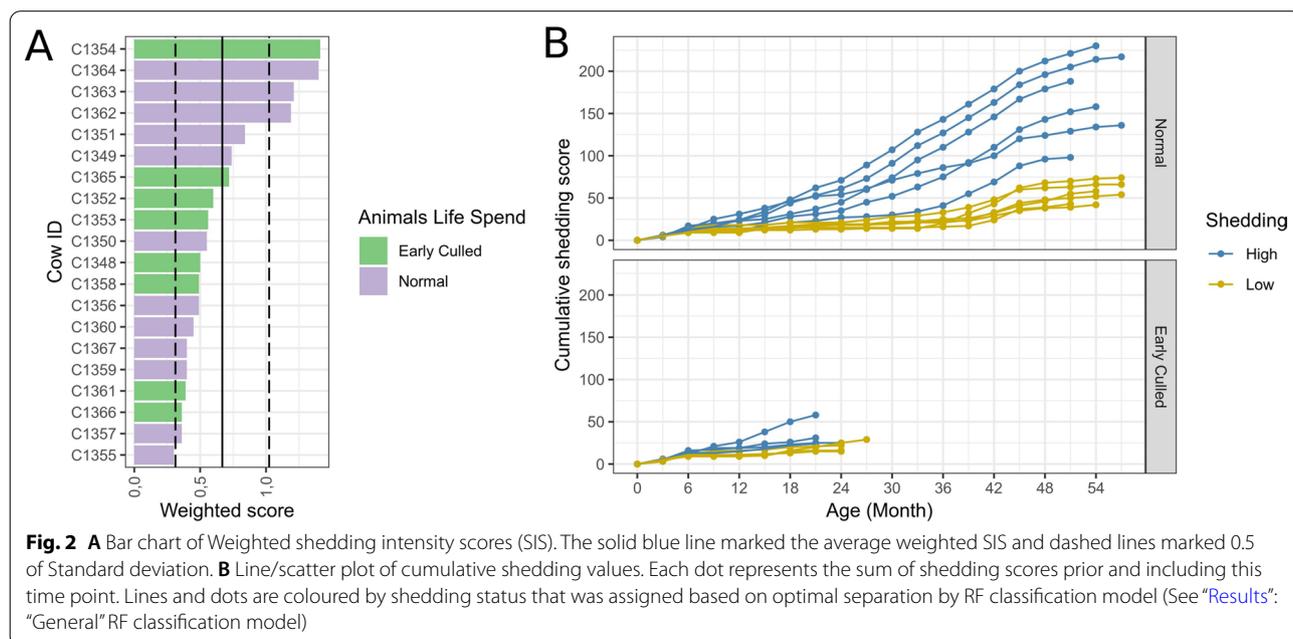
Samples sequencing and overall microbial composition

Two hundred fifty seven samples remained in the sample-set after the quality control with a median of 12 samples per cow, a maximum of 22, and a minimum of 8 (Additional file 4: Table S1). In total, 3,973,061 reads passed the initial quality control and filtering as described in the material and methods section. The median value of reads per sample was 14,000 reads with a minimum of 1415 and a maximum of 44,407 (Additional file 4: Fig. S2A). The rarefaction curves showed sufficient sequencing depth to capture microbial diversity at ASV level (Additional file 4: Fig. S2 B).

Overall 4940 ASVs from 15 microbial phyla remained in the dataset after the filtering steps. However, only *Firmicutes* (from 61.3 to 64.3%, 95% CI), *Bacteroidota* (from 29.2 to 32.2%, 95% CI), *Verrucomicrobiota* (from 2.0 to 2.7%, 95% CI), *Actinobacteriota* (from 1.3 to 1.7%, 95% CI) received averagely more than 1% of the reads across all samples (Additional file 4: Fig. S3, Table S2).

Prediction of weighted SISs by RF regression model based on microbiota composition

The absolute differences between predicted and actual Weighted SISs ranged from 0.0007 up to 0.81, with the median value of 0.2. A significant correlation between animal ID and absolute difference between predicted and actual Weighted SISs (Pearson; $r = -0.85$, $p < 0.0001$) was found, but not with the age of the animals. Overall averages of the predicted Weighted SISs were significantly correlated with the actual Weighted SISs (Pearson; $r = 0.83$, $p < 0.0001$), however, accuracy of the prediction



varied strongly per individual sample and per animal (Additional file 4: Fig. S4).

“General” RF classification model

The best prediction outcome for the RF model based on all ASVs (“General” RF) was observed when samples with the Weighted SISs ≤ 0.5 were assigned as “Low” shedders and all above assigned as “High”. At that point, class errors for the “Low” and “High” shedders were comparable with each other (Fig. 3).

A larger number of trees (ntree = 15,001) resulted in a more stable classification model, and the accuracy of classification increased when a larger than the default number of features (ASVs) per split (mtry = 568; Additional file 4: Fig. S5) was used. The final classification model resulted in an OOB error of 22.2% and class errors of 28.7% and 16.3% for “High” and “Low” shedding groups respectively.

ASVs significantly contributing to classification

In total, 192 ASVs were identified as significantly contributing to the accuracy of classification (Additional file 5). From the phylogenetic point of view, identified ASVs were distributed among nine different phyla, 12 classes, 25 orders, 41 families and 54 genera.

The majority of ASVs that contributed to classification accuracy belonged to *Firmicutes* (108 ASVs) and *Bacteroidota* (58 ASVs) phyla followed by *Verrucomicrobiota* (10 ASVs) and *Actinobacteriota* (7 ASVs), and the remaining phyla were represented only by 1–3 ASVs. The overall composition of the significantly contributing taxa mirrored the overall microbiota composition, however, out of the top 21 ASVs contributing

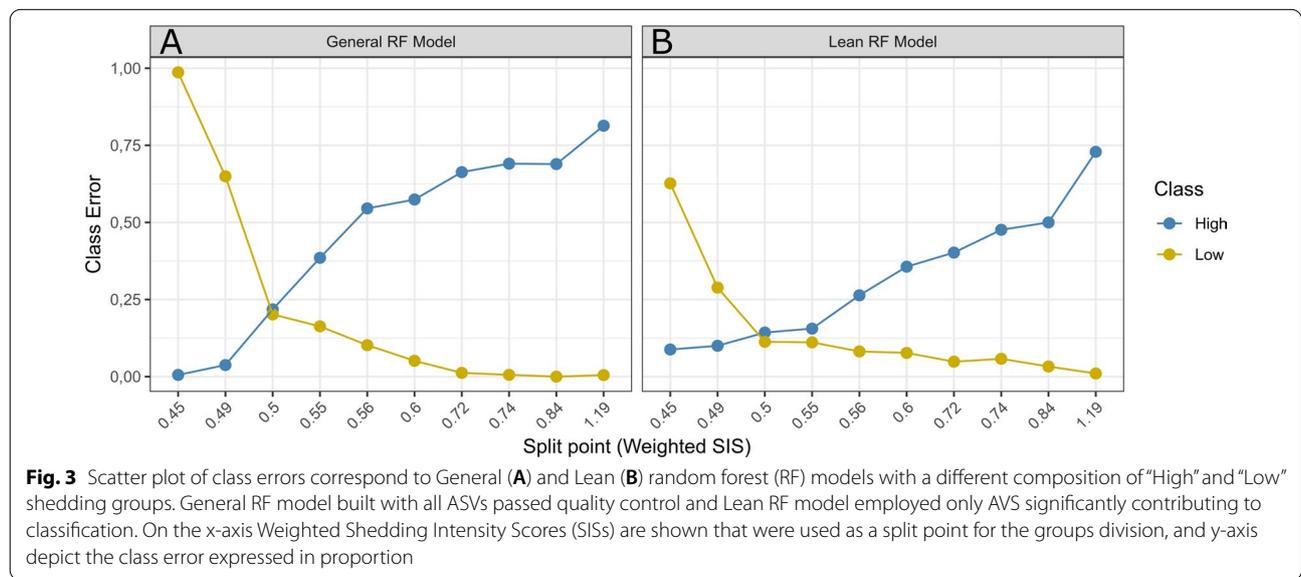
to the Accuracy of the RF classification model, more ASVs belonged to *Bacteroidota* (13 ASVs) than to *Firmicutes* (6 ASVs; Fig. 4).

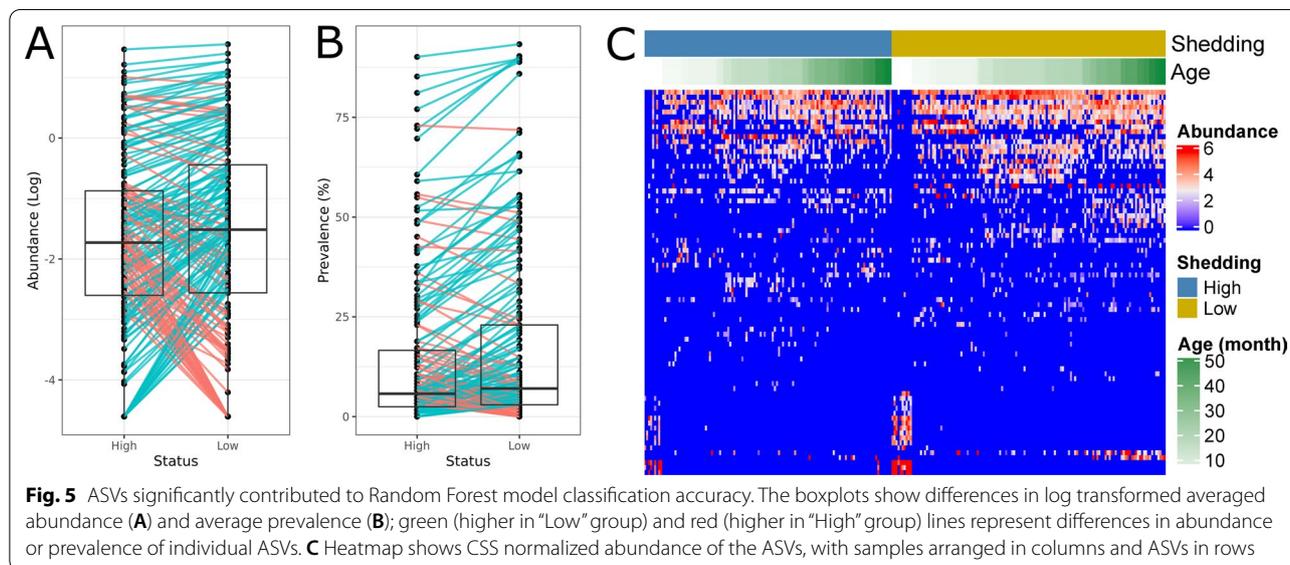
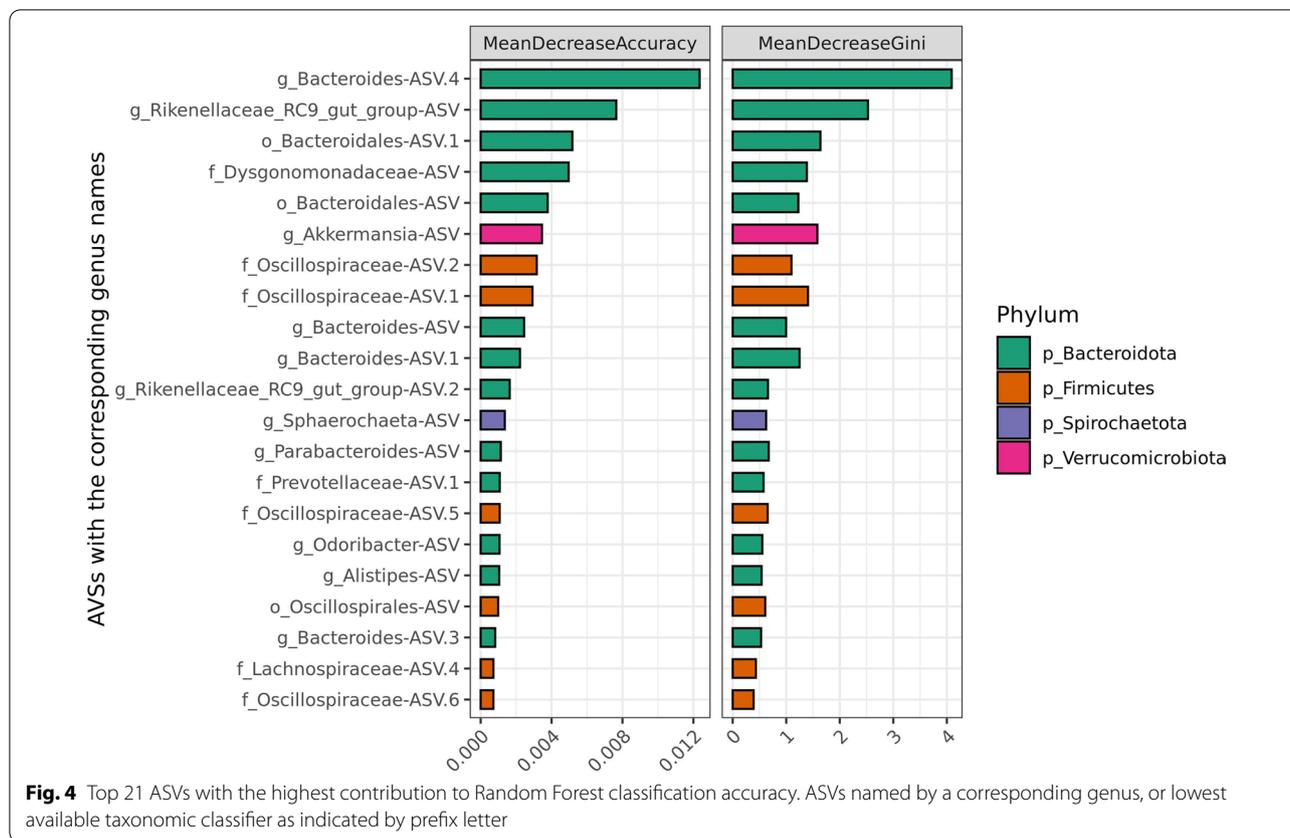
The differences in the average CSS normalized abundance and the prevalence of ASVs contributed to the classification accuracy between “High” and “Low” shedding groups ranged from 0.01 to 1.59 and from 0.5 to 34% respectively. Overall, no large difference in ASVs prevalence, nor abundance in the “Low” versus “High” shedding groups (Fig. 5A, B) was observed. However, a clear distinction in ASVs identified as significantly contributing to classification was observed between samples from the animals with the milk (pre-weaning) and solid diet (post-weaning; Fig. 5C).

“Lean” RF classification model

The RF classification model was built using only ASVs that were identified as significantly contributing to classification (“Lean” RF classification model) with the goal to reduce the number of input features, decrease computational time and subsequently improve the accuracy of classification. Similarly to the “General” RF model, the best split for the “Lean” RF model was observed when samples with the Weighted SIS ≤ 0.5 were assigned to the “Low” shedding group, and all above to the “High” shedding group. However, the “Lean” RF model showed a similar classification accuracy for the Weighted SIS split point 0.5 and 0.55, rather than a single best split point as it was observed for the General RF model (Fig. 3).

The “Lean” RF model showed the best accuracy of classification when 28 ASVs per split (mtry parameter) and an excessively sufficient number of trees (ntree = 15,001) were used. The “Lean” RF model showed higher accuracy

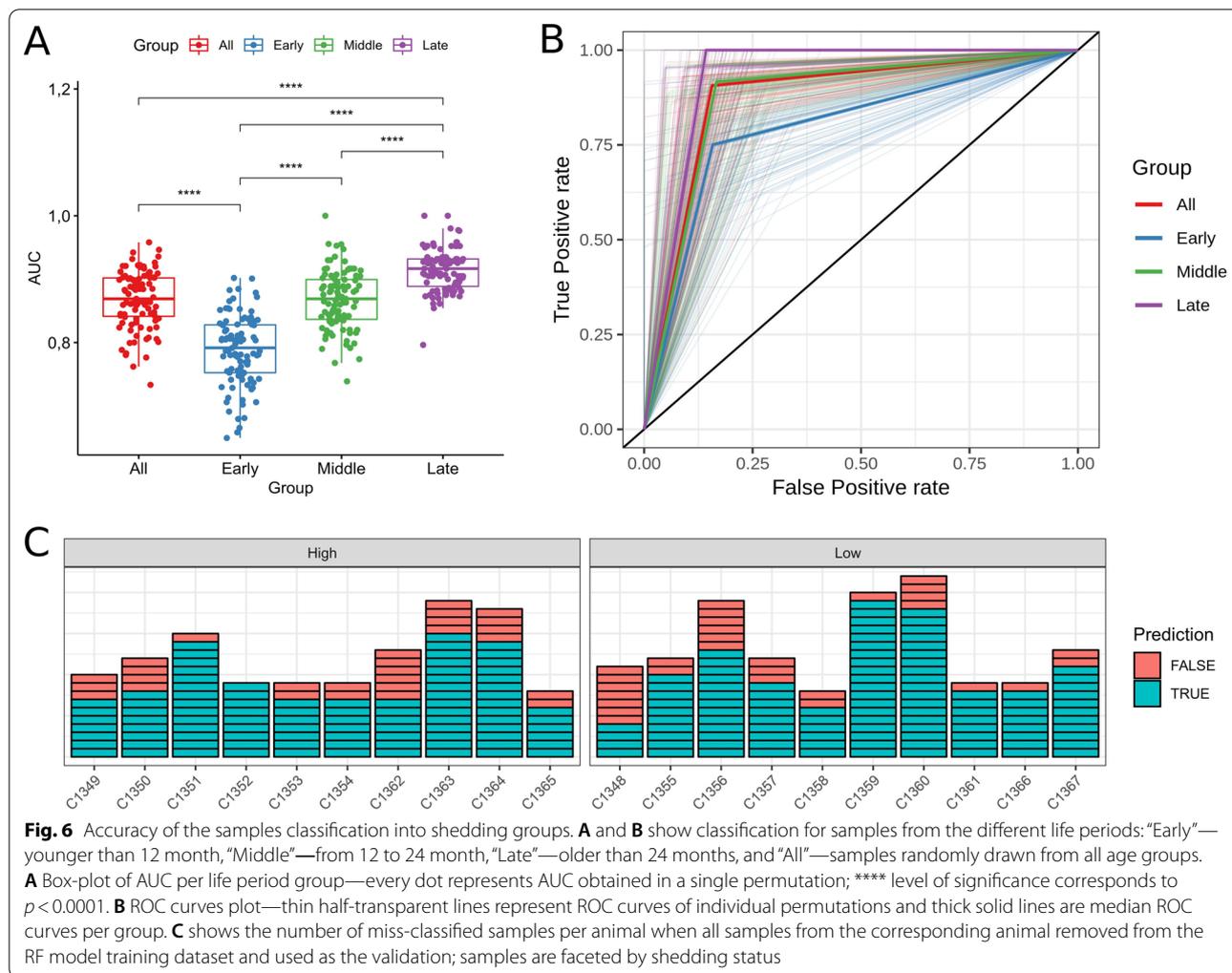




than the “General” RF model and resulted in 12.84% OOB. The class errors were 9.63% for the “Low” and 16.4% for the “High” shedding groups.

Influence of age on the accuracy of the shedding status prediction

A progressive improvement in the accuracy of the shedding status prediction from the “Early” to “Late” age groups was observed (Fig. 6A, B). The highest AUC was



observed for the samples from the animals older than 24 months (“Late” group; AUC from 0.91 to 0.92, 95% CI) and the lowest in the samples from the animals younger than 12 months (“Early” group; AUC from 0.78 to 0.79, 95% CI). Prediction for samples from animals between 12 and 24 month of age (“Middle” group) showed an intermediate accuracy (AUC from 0.86 to 0.87, 95% CI) that was not significantly different from the accuracy observed for the complete dataset (“All” group; AUC from 0.86 to 0.87, 95% CI). In addition, a higher consistency in observer AUC between permutations was observed in the “Late” (SD = 0.036) group in comparison with “Middle” (SD = 0.046) and “Early” (SD = 0.054) groups.

Prediction of the shedding status controlled for microbiota similarity within an animal

The number of correctly classified samples varied per animal (Fig. 6C) with 95% CI between 70.3% and 83.7%. However, only 5 out of 11 samples from cow C1348 were

classified correctly. The overall AUC with this classification approach was 0.77. No clear age dependent pattern for misclassification of the samples was observed (Additional file 4: Fig. S6).

Discussion

To our knowledge, this is the first study that leveraged relationships between the gut microbiota composition and MAP infection in cows to predict severity of MAP shedding in the environment. Effective diagnosis of contagious diseases is a cornerstone when it comes to their control and prevention. However, a positive test does not always provide guidance for further action, where a forecasting model could give a direction for disease control options and mitigation of financial losses. It is particularly relevant for diseases with a slow progressive development and a variable outcome such as Johne’s disease.

Our prediction model in the best-case scenario showed an AUC from 0.86 to 0.87 (95% CI) for

separation of the “Low” from the “High” shedding cows. It is difficult to compare these results with the conventional diagnostic methods based on culturing, PCR or ELISA. The conventional MAP diagnostic methods are designed to give the information about presence or absence of the pathogen at the time of sampling rather than predict development of the disease and amount of the MAP shed by the infected cow during its lifetime.

Accuracy of our model was age sensitive and showed improvement from young to the old age groups. Similar observations have been made for conventional diagnostic methods. The specificity of culture and ELISA based methods has been shown to be greatly improved in older cows with more advanced stages of the disease [39, 40]. The slow-progressive nature of MAP infection, especially the slow development of lesions as well as MAP specific immune responses is the most likely to be the explanation for these observations. Improved sensitivity of the microbiota based model could be attributed to extensive pathogen-induced changes in the intestinal environment and the microbiome, where the sensitivity of culturing or PCR based methods solely relies on the increase of MAP replication and faecal shedding. This difference is important, since decoupling of a predictive or diagnostic method from the presence of the pathogen in a sample could greatly improve robustness of the method. MAP cells tend to clump together rather than grow as a planktonic culture, which inherently creates an uneven distribution of MAP in faecal samples. The heterogeneous distribution of MAP in biological matrices can seriously compromise sensitivity of tests that are reliant on the presence of MAP in a small tested sample. Currently, only immunological tests (predominantly absorbed ELISA tests) are decoupled from the direct detection of MAP in samples. However, immunological tests also struggle with ambiguous results due to arbitrary defined cut-off values between positive and negative samples.

Detection of MAP in early stages of the disease with classical methods is particularly difficult due to intermittent or lack of pathogen shedding [41] as well as slow development of the adaptive immune response [42]. The results of the current study however show satisfactory accuracy (AUC from 0.78 to 0.79, 95% CI) of classification into “Low” and “High” shedding groups even at an early stage of infection. Based on these results it is tempting to speculate about the possibility of microbiota composition based diagnostic tests that will be superior in detection of the paratuberculosis at the very early stages of the disease. However, that will require an experiment with a different study design that will allow the comparison between a large number of healthy and cows infected with MAP.

The microbiota based shedding intensity classifier was built using the RF algorithm. The RF shown to be a simple, effective and “white box” machine learning approach that can handle various data types [14], including microbiota compositional data [43, 44]. Flexibility of the approach allowed construction of an accurate classifier, but also identification of the MAP infection relevant ASVs biomarkers and test possibilities.

Construction of the “General” RF model allowed the identification of ASVs for the final (“Lean”) classification model. When comparing performance of the “General” and the “Lean” RF model a clear improvement in accuracy of classification as well as a reduction of the time required for the model training was observed. The selection of features is commonly used prior to application of a machine learning algorithm [45] to reduce the amount of noise from irrelevant data, and overcome the curse of dimensionality [46].

The RF regression model that was developed showed a moderate performance. Despite a statistically significant correlation between predicted and actual Weighted SISs the accuracy was not satisfactory. In particular, the samples for the animals with lower Weighted SISs had a large discrepancy between the actual and predicted values. Nevertheless, a regression model to predict shedding of MAP could be a useful tool in certain situations, but the model construction will require a larger dataset and a different experimental design.

Optimization of RF parameters was an important step to improve accuracy of the classification model. This sensitivity of RF to changes in *mtry* parameter was shown previously in application to the gene expression data [47].

Definition of a response variable and groups for classification is a crucially important step in development of a the classification model. In our case, the shedding intensity data was inferred from the number of cultured MAP colonies and was used as the response variable. The scores of the shedding intensity rather than actual count of the cultivated colonies were used to allow a higher level of generalization, which could be beneficial for smaller data-sets. Nevertheless, the actual count of cultivated colonies or qPCR assessed gene copy number could be useful for larger datasets, and could allow for inference of the shedding values rather than classification into shedding intensity groups. However, due to high variability of microbiota composition between individual animals this approach could also suffer from a low precision with less interpretable results.

When it comes to the definition of the groups for classification—there is no clear guide what could be considered low or high shedding cow, and cut-off is usually set arbitrarily. A gradual increase in the Weighted SISs in studied animals was observed without a clear separation

into “High” or “Low” shedders. MAP infected animals often show intermediate shedding where periods of severe shedding are followed by low or completely shedding free periods [41, 48]. Therefore, when samples are not collected longitudinally the division into high and low shedders could be more prominent [49]. However, differences became more subtle when shedding score was evaluated during the lifetime of the animal. An optimal separation into “Low” and “High” shedders was obtained by comparing the accuracy of RF models built with the different configurations of the groups. This approach is suitable for the “proof of concept” research, such as presented in this paper, however, the development of a real world applicable MAP shedding prediction tools will require a larger number of animals with a clearer division into shedding groups.

One of the great advantages of the RF classification algorithm is the possibility to get insight in the contribution of individual features to classification. In total, 4940 ASVs were identified across all samples, nevertheless, only 4% from the total number was important for the classification. This is not surprising since not every microbial taxa will respond to MAP infection, or have protective qualities against it. In addition, microbiota composition has been shown to vary between individuals and only a relatively small number of taxa comprises a shared core [50, 51]. The current dataset consisted from samples collected in a longitudinal experiment that captured all stages of animal development with the consequent age and season related microbiota variations, which further decreased the number of ASVs that could have a meaningful impact on the classification. In the current study, age related change in diet from milk to solid food and following overall microbial composition are reflected in differences in ASVs significant to classification. This is an expected finding, nevertheless, it underlines the importance of an experimental design adequate to reach the intended goals. In our case, the number of samples available from the animals of early age (milk diet) was a small proportion of total samples and did not allow for a separate investigation.

Among the top 20 ASVs contributing to classification, the majority had a significantly higher abundance in the “Low” (18 ASVs) than in the “High” (2 ASVs) shedding group. It is difficult to speculate about the function of ASVs and why they are more abundant in the “Low” shedding group, due to the limited taxonomic resolution of the microbiota profiling based on an amplified 16S rRNA fragment [52]. In addition, almost half (8 out of 20) ASVs were not assigned even to a genus level, increasing uncertainty regarding the role of a microbial feature. Future research may benefit from advances in sequence technologies and strategies such as shotgun metagenomic

sequencing. Nevertheless, even a high-level taxonomic assignment can give an indication of the ecological niche, albeit with limitations. ASVs from genus *Bacteroides* showed the highest contribution to the accuracy of classification. Four out of the top 20 ASVs belonged to genus *Bacteroides*, and all of them have significantly higher abundance in the “Low” in comparison with the “High” shedding group. Members of the genus *Bacteroides* can be found only as a part of the mammalian gastrointestinal microbiota [53]. They are generally regarded as beneficial commensals when they resign in their niche, however, some strains are pathogenic and can cause severe infections [54]. Higher abundance of *Bacteroides* in the “Low” shedding group could be an indication of their protective properties and competitive exclusion of MAP, however further research will be needed. Intriguingly, an ASV from the genus *Akkermansia* had a significantly higher abundance in the “High” shedding group. *Akkermansia* is a mucin degrading commensal bacteria with a tight connection with the host [55]. It is generally considered as a beneficial microbe in humans [56]. The higher abundance in the “High” shedding group is puzzling, but could be explained by physiological changes in the mucosal layer due to MAP colonization and as a consequence an expansion of the *Akkermansia* habitat. However, the possibility of a more direct relationship between MAP and *Akkermansia* should not be discarded.

Conclusions

In this proof-of-concept-study, showed evidence that future shedding severity of MAP by cows can potentially be forecast based on the composition of the intestinal microbiota. To our knowledge this is the first approach that showed a potential to predict severity of MAP infection at early stages of the infection. Forecasting of MAP infection progression will help to develop management strategies in farms with the high prevalence, endemic and or hard to control MAP infection as it will provide farmers and veterinarians with a tool to select animals for culling. An indiscriminate culling based on MAP specific test results indicating presence or absence of MAP may be an economically non-viable strategy for areas with endemic MAP. Therefore, microbiota based prediction of the infection development can help mitigate the economic burden of MAP and limit spread of the infection in populations where MAP is endemic by preferentially targeting animals with a likely higher lifetime contribution to MAP transmission and environmental contamination.

Abbreviations

MAP: *Mycobacterium Avium* Subsp. *Paratuberculosis*; RF: Random forest; OOB: Out of bag error; ASV: Amplicon sequence variant; PCR: Polymerase chain reaction; ELISA: Enzyme-linked immunosorbent assay; AUC: Area under the

curve; SIS: Shedding Intensity Score; CI: Confidence interval; ROC: Receiver operating characteristic; CSS: Cumulative sum scaling.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s42523-021-00143-y>.

Additional file 1: Sample selection data.

Additional file 2: Sample metadata.

Additional file 3: Fecal shedding intensity data.

Additional file 4: Shedding intensity score calculation.

Additional file 5: Significantly contributing ASV.

Acknowledgements

Not applicable.

Authors' contributions

AU developed the study concept, analysed and interpreted the data and wrote the draft manuscript. AD, SV and LR handled faecal samples, extracted DNA and prepared samples for sequencing. AK participated in the study concept development, data interpretation and the manuscript writing. All authors participated in the manuscript writing and revision.

Funding

This research received financial support from the Dutch Ministry of Agriculture, Nature and Food Quality (WOT-01-002-002.02).

Availability of data and materials

The raw sequencing data generated during the current study are available in the NCBI repository under BioProject No. PRJNA706519. Data will be accessible upon publication. Metadata provided as an additional file with the manuscript. Full data analysis pipeline is available as a GitHub repository via the link: https://github.com/AlexanderUm/WBVR_MAP_Microbiota.

Declarations

Ethics approval and consent to participate

The animal experiments described in this study were performed in strict accordance with the provisions of the European Convention for the protection of vertebrate animals used for experimental and other scientific purposes (86/609 EG). The animal experiments were approved by the Animal Welfare Body of WBVR permit number 299-47053-07/99-01 in accordance with the Dutch regulations on animal experimentation.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Department of Bacteriology, Host Pathogen Interaction and Diagnostics Development, Wageningen Bioveterinary Research, Lelystad, The Netherlands. ²Department of Population Health Sciences, Faculty of Veterinary Medicine, Utrecht University, Utrecht, The Netherlands. ³Present Address: Chair Group Youth Food and Health, Faculty of Science and Engineering, Maastricht University Campus Venlo, Venlo, The Netherlands.

Received: 12 July 2021 Accepted: 31 October 2021

Published online: 14 November 2021

References

- Cocito C, Gilot P, Coene M, de Kesel M, Poupart P, Vannuffel P. Paratuberculosis. *Clin Microbiol Rev.* 1994;7(3):328–45. <https://doi.org/10.1128/cmr.7.3.328>.
- Sweeney RW. Transmission of paratuberculosis. *Vet Clin N Am Food Anim Pract.* 1996;12(2):305–12. [https://doi.org/10.1016/S0749-0720\(15\)30408-4](https://doi.org/10.1016/S0749-0720(15)30408-4).
- Taniguchi Y, Sakakibara S-I, Fujihara M, Yagi A, Fujiyoshi S. The association between detection of *Mycobacterium avium* subsp. *paratuberculosis* DNA in feces and histopathological classification. *J Vet Med Sci.* 2020;82(5):541–5. <https://doi.org/10.1292/jvms.18-0724>.
- Biemans F, Ben Romdhane R, Gontier P, Fourichon C, Ramsbottom G, More SJ, et al. Modelling transmission and control of *Mycobacterium avium* subspecies *paratuberculosis* within Irish dairy herds with compact spring calving. *Prev Vet Med.* 2021;186:105228. <https://doi.org/10.1016/j.prevetmed.2020.105228>.
- Whitlock RH, Wells SJ, Sweeney RW, Van Tiem J. ELISA and fecal culture for paratuberculosis (Johne's disease): sensitivity and specificity of each method. *Vet Microbiol.* 2000;77(3):387–98. [https://doi.org/10.1016/S0378-1135\(00\)00324-2](https://doi.org/10.1016/S0378-1135(00)00324-2).
- Nielsen SS, Grønbaek C, Agger JF, Houe H. Maximum-likelihood estimation of sensitivity and specificity of ELISAs and faecal culture for diagnosis of paratuberculosis. *Prev Vet Med.* 2002;53(3):191–204.
- Koets AP, Eda S, Sreevatsan S. The within host dynamics of *Mycobacterium avium* ssp. *paratuberculosis* infection in cattle: where time and place matter. *Vet Res.* 2015;46(1):61. <https://doi.org/10.1186/s13567-015-0185-0>.
- Bannantine JP, Bermudez LE. No holes barred: invasion of the intestinal mucosa by *Mycobacterium avium* subsp. *paratuberculosis*. *Infect Immun.* 2013;81(11):3960. <https://doi.org/10.1128/IAI.00575-13>.
- Derakhshani H, De Buck J, Mortier R, Barkema HW, Krause DO, Khafipour E. The features of fecal and ileal mucosa-associated microbiota in dairy calves during early infection with *Mycobacterium avium* Subspecies *paratuberculosis*. *Front Microbiol.* 2016. <https://doi.org/10.3389/fmicb.2016.00426>.
- Fecteau M-E, Pitta DW, Vecchiarelli B, Indugu N, Kumar S, Gallagher SC, et al. Dysbiosis of the fecal microbiota in cattle infected with *Mycobacterium avium* subsp. *paratuberculosis*. *PLoS ONE.* 2016;11(8):e0160353. <https://doi.org/10.1371/journal.pone.0160353>.
- Arrazuria R, Elguezal N, Juste RA, Derakhshani H, Khafipour E. *Mycobacterium avium* Subspecies *paratuberculosis* infection modifies gut microbiota under different dietary conditions in a rabbit model. *Front Microbiol.* 2016. <https://doi.org/10.3389/fmicb.2016.00446>.
- Adlung L, Elinav E, Greten TF, Korangy F. Microbiome genomics for cancer prediction. *Nat Cancer.* 2020;1(4):379–81.
- Liu Y, Meric G, Havulinna AS, Teo SM, Ruuskanen M, Sanders J, et al. Early prediction of liver disease using conventional risk factors and gut microbiome-augmented gradient boosting. *medRxiv.* 2020. <https://doi.org/10.1101/2020.06.24.20138933>.
- Olivares M, Walker AW, Capilla A, Benítez-Páez A, Palau F, Parkhill J, et al. Gut microbiota trajectory in early life may predict development of celiac disease. *Microbiome.* 2018;6(1):36.
- Zhou Y, Xu ZZ, He Y, Yang Y, Liu L, Lin Q, et al. Gut microbiota offers universal biomarkers across ethnicity in inflammatory bowel disease diagnosis and infliximab response prediction. *mSystems.* 2018;3(1):e00188-17. <https://doi.org/10.1128/mSystems.00188-17>.
- Kalliomäki M, Carmen Collado M, Salminen S, Isolauri E. Early differences in fecal microbiota composition in children may predict overweight. *Am J Clin Nutr.* 2008;87(3):534–8.
- Faith JJ, Guruge JL, Charbonneau M, Subramanian S, Seedorf H, Goodman AL, et al. The long-term stability of the human gut microbiota. *Science.* 2013;341(6141):1237439. <https://doi.org/10.1126/science.1237439>.
- Huang S, Ji S, Wang F, Huang J, Alugongo GM, Li S. Dynamic changes of the fecal bacterial community in dairy cows during early lactation. *AMB Express.* 2020. <https://doi.org/10.1186/s13568-020-01106-3>.
- Tang MT, Han H, Yu Z, Tsuruta T, Nishino N. Variability, stability, and resilience of fecal microbiota in dairy cows fed whole crop corn silage. *Appl Microbiol Biotechnol.* 2017;101(16):6355–64.
- Zhang Q, Li C, Niu X, Zhang Z, Li F, Li F. The effects of milk replacer allowance and weaning age on the performance, nutrients digestibility, and ruminal microbiota communities of lambs. *Anim Feed Sci Technol.* 2019;257:114263.

21. Odamaki T, Kato K, Sugahara H, Hashikura N, Takahashi S, Xiao J-Z, et al. Age-related changes in gut microbiota composition from newborn to centenarian: a cross-sectional study. *BMC Microbiol.* 2016;16(1):90. <https://doi.org/10.1186/s12866-016-0708-5>.
22. Midani FS, Weil AA, Chowdhury F, Begum YA, Khan AI, Debela MD, et al. Human gut microbiota predicts susceptibility to vibrio cholerae infection. *J Infect Dis.* 2018;218(4):645–53. <https://doi.org/10.1093/infdis/jiy192>.
23. Shalev-Shwartz S, Ben-David S. Understanding machine learning: from theory to algorithms. Cambridge: Cambridge University Press; 2014.
24. Marcos-Zambrano LJ, Karadzovic-Hadziabdic K, Loncar Turukalo T, Przymus P, Trajkovic V, Aasmets O, et al. Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction and treatment. *Front Microbiol.* 2021. <https://doi.org/10.3389/fmicb.2021.634511>.
25. Langelaar MFM. Heat shock protein 70 and bovine paratuberculosis. Utrecht: Utrecht University; 2005.
26. Ganusov VV, Klinkenberg D, Bakker D, Koets AP. Evaluating contribution of the cellular and humoral immune responses to the control of shedding of *Mycobacterium avium* spp. *paratuberculosis* in cattle. *Vet Res.* 2015;46(1):62.
27. Vary PH, Andersen PR, Green E, Hermon-Taylor J, McFadden JJ. Use of highly specific DNA probes and the polymerase chain reaction to detect *Mycobacterium paratuberculosis* in Johnne disease. *J Clin Microbiol.* 1990;28(5):933.
28. Herlemann DPR, Labrenz M, Jürgens K, Bertilsson S, Waniek JJ, Andersson AF. Transitions in bacterial communities along the 2000 km salinity gradient of the Baltic Sea. *ISME J.* 2011;5(10):1571–9.
29. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: high-resolution sample inference from Illumina amplicon data. *Nat Methods.* 2016;13(7):581–3.
30. Team RC. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2017.
31. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 2013;41(D1):D590–6. <https://doi.org/10.1093/nar/gks1219>.
32. McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE.* 2013;8(4):e61217.
33. Paulson JN, Stine OC, Bravo HC, Pop M. Differential abundance analysis for microbial marker-gene surveys. *Nat Methods.* 2013;10(12):1200–2.
34. Breiman L. Random forests. *Mach Learn.* 2001;45(1):5–32.
35. Liaw A, Wiener M. Classification and regression by RandomForest. *Forest.* 2001;23.
36. Archer E. rFPermute: estimate permutation p-values for Random Forest importance metrics. R package version. 2016;1(2).
37. Wickham H. The tidyverse. R package ver. 2017;1(1).
38. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics.* 2016;32(18):2847–9.
39. Nielsen SS, Toft N. Age-specific characteristics of ELISA and fecal culture for purpose-specific testing for paratuberculosis. *J Dairy Sci.* 2006;89(2):569–79. [https://doi.org/10.3168/jds.S0022-0302\(06\)72120-8](https://doi.org/10.3168/jds.S0022-0302(06)72120-8).
40. Blanco Vázquez C, Alonso-Hearn M, Juste RA, Canive M, Iglesias T, Iglesias N, et al. Detection of latent forms of *Mycobacterium avium* subsp. *paratuberculosis* infection using host biomarker-based ELISAs greatly improves paratuberculosis diagnostic sensitivity. *PLoS ONE.* 2020;15(9):e0236336.
41. Mortier RAR, Barkema HW, Orsel K, Wolf R, De Buck J. Shedding patterns of dairy calves experimentally infected with *Mycobacterium avium* subspecies *paratuberculosis*. *Vet Res.* 2014;45(1):71. <https://doi.org/10.1186/s13567-014-0071-1>.
42. Huda A, Jungersen G, Lind P. Longitudinal study of interferon-gamma, serum antibody and milk antibody responses in cattle infected with *Mycobacterium avium* subsp. *paratuberculosis*. *Vet Microbiol.* 2004;104(1):43–53. <https://doi.org/10.1016/j.vetmic.2004.08.011>.
43. Belk A, Xu ZZ, Carter DO, Lynne A, Bucheli S, Knight R, et al. Microbiome data accurately predicts the postmortem interval using random forest regression models. *Genes.* 2018. <https://doi.org/10.3390/genes9020104>.
44. Saulnier DM, Riehle K, Mistretta TA, Diaz MA, Mandal D, Raza S, et al. Gastrointestinal microbiome signatures of pediatric patients with irritable bowel syndrome. *Gastroenterology.* 2011;141(5):1782–91. <https://doi.org/10.1053/j.gastro.2011.06.072>.
45. Blum AL, Langley P. Selection of relevant features and examples in machine learning. *Artif Intell.* 1997;97(1):245–71. [https://doi.org/10.1016/S0004-3702\(97\)00063-5](https://doi.org/10.1016/S0004-3702(97)00063-5).
46. Qi Y. Random forest for bioinformatics. In: Ensemble machine learning. Springer; 2012. p. 307–23.
47. Statnikov A, Wang L, Aliferis CF. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinform.* 2008;9(1):319. <https://doi.org/10.1186/1471-2105-9-319>.
48. Raizman EA, Fetrow J, Wells SJ, Godden SM, Oakes MJ, Vazquez G. The association between *Mycobacterium avium* subsp. *paratuberculosis* fecal shedding or clinical Johnne's disease and lactation performance on two Minnesota, USA dairy farms. *Prevent Vet Med.* 2007;78(3):179–95. <https://doi.org/10.1016/j.prevetmed.2006.10.006>.
49. Crossley BM, Zagmutt-Vergara FJ, Fyock TL, Whitlock RH, Gardner IA. Fecal shedding of *Mycobacterium avium* subsp. *paratuberculosis* by dairy cows. *Vet Microbiol.* 2005;107(3):257–63. <https://doi.org/10.1016/j.vetmic.2005.01.017>.
50. Durso LM, Harhay GP, Smith TPL, Bono JL, DeSantis TZ, Harhay DM, et al. Animal-to-animal variation in fecal microbial diversity among beef cattle. *Appl Environ Microbiol.* 2010;76(14):4858. <https://doi.org/10.1128/AEM.00207-10>.
51. Falony G, Joossens M, Vieira-Silva S, Wang J, Darzi Y, Faust K, et al. Population-level analysis of gut microbiome variation. *Science.* 2016;352(6285):560. <https://doi.org/10.1126/science.aad3503>.
52. Fuks G, Elgart M, Amir A, Zeisel A, Turnbaugh PJ, Soen Y, et al. Combining 16S rRNA gene variable regions enables high-resolution microbial community profiling. *Microbiome.* 2018;6(1):17. <https://doi.org/10.1186/s40168-017-0396-x>.
53. Ley RE, Hamady M, Lozupone C, Turnbaugh PJ, Ramey RR, Birchler JS, et al. Evolution of mammals and their gut microbes. *Science.* 2008;320(5883):1647. <https://doi.org/10.1126/science.1155725>.
54. Wexler HM. Bacteroides: the good, the bad, and the nitty-gritty. *Clin Microbiol Rev.* 2007;20(4):593. <https://doi.org/10.1128/CMR.00008-07>.
55. Derrien M, Vaughan EE, Plugge CM, de Vos WM. Akkermansia muciniphila gen. nov., sp. nov., a human intestinal mucin-degrading bacterium. *Int J Syst Evolut Microbiol.* 2004;54(5):1469–76.
56. Belzer C, De Vos WM. Microbes inside—from diversity to function: the case of Akkermansia. *ISME J.* 2012;6(8):1449–58.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.