Journal of Breath Research

PAPER • OPEN ACCESS

Strategies for the identification of disease-related patterns of volatile organic compounds: prediction of paratuberculosis in an animal model using random forests

To cite this article: Elisa Kasbohm *et al* 2017 *J. Breath Res.* **11** 047105

View the article online for updates and enhancements.

Related content

- Physiological variability in volatile organic compounds (VOCs) in exhaled breath and released from faeces due to nutrition and somatic growth in a standardized caprine animal model4
Sina Fischer, Phillip Trefz, Andreas Bergmann et al.
- Effects of biological and methodological factors on volatile organic compound patterns during cultural growth of Mycobacterium avium ssp. paratuberculosis
Anne Küntzel, Sina Fischer, Andreas Bergmann et al.
- The versatile use of exhaled volatile organic compounds in human health and disease
Agnes W Boots, Joep J B N van Berkel, Jan W Dallinga et al.

This content was downloaded from IP address 193.22.115.2 on 02/11/2017 at 06:47

Journal of Breath Research

OPEN ACCESS

RECEIVED
15 May 2017

REVISED
21 July 2017

ACCEPTED FOR PUBLICATION
3 August 2017

PUBLISHED
1 November 2017

Original content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence.

Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

PAPER

Strategies for the identification of disease-related patterns of volatile organic compounds: prediction of paratuberculosis in an animal model using random forests

Elisa Kasbohm[1,2], Sina Fischer[3], Anne Küntzel[3], Peter Oertel[4], Andreas Bergmann[4], Phillip Trefz[4], Wolfram Miekisch[4], Jochen K Schubert[4], Petra Reinhold[3], Mario Ziller[5], Andreas Fröhlich[1], Volkmar Liebscher[2] and Heike Köhler[3]

[1] Institute of Epidemiology, Friedrich-Loeffler-Institut (FLI), Federal Research Institute for Animal Health, Greifswald—Insel Riems, Germany
[2] Department of Mathematics and Computer Science, University of Greifswald, Germany
[3] Institute of Molecular Pathogenesis, Friedrich-Loeffler-Institut (FLI), Federal Research Institute for Animal Health, Jena, Germany
[4] Rostock Medical Breath Research Analytics and Technologies (RoMBAT), Department of Anesthesia and Intensive Care, Rostock University Medical Center, Germany
[5] Workgroup Biomathematics, Friedrich-Loeffler-Institut (FLI), Federal Research Institute for Animal Health, Greifswald—Insel Riems, Germany

E-mail: heike.koehler@fli.de

Keywords: biomarker discovery, effect size, exhaled breath, faeces, paratuberculosis, random forest, volatile organic compound (VOC)

Supplementary material for this article is available online

Abstract

Modern statistical methods which were developed for pattern recognition are increasingly being used for data analysis in studies on emissions of volatile organic compounds (VOCs). With the detection of disease-related VOC profiles, novel non-invasive diagnostic tools could be developed for clinical applications. However, it is important to bear in mind that not all statistical methods are equally suitable for the investigation of VOC profiles. In particular, univariate methods are not able to discover VOC patterns as they consider each compound separately. The present study demonstrates this fact in practice. Using VOC samples from a controlled animal study on paratuberculosis, the random forest classification method was applied for pattern recognition and disease prediction. This strategy was compared with a prediction approach based on single compounds. Both methods were framed within a cross-validation procedure. A comparison of both strategies based on these VOC data reveals that random forests achieves higher sensitivities and specificities than predictions based on single compounds. Therefore, it will most likely be more fruitful to further investigate VOC patterns instead of single biomarkers for paratuberculosis. All methods used are thoroughly explained to aid the transfer to other data analyses.

1. Introduction

The potential of volatile organic compounds (VOCs) for diagnostic purposes has already been acknowledged for a variety of diseases [1, 2]. Since VOCs are emitted constantly during various metabolic processes, the detection of disease-related VOC profiles might enable the development of novel non-invasive diagnostic tools [3]. Nevertheless, the high naturally occurring variability which is observed in measurements on VOC samples currently hinders potential clinical applications [4]. This variability originates to a large extent in external confounding factors and physiological effects [5, 6]. At the same time, recent techniques in metabolomics (like mass spectrometry and nuclear magnetic resonance) enable hundreds of compounds to be measured for each sample. Therefore, data analysis is central in order to assess if observed variations are related to a disease and to identify potential biomarkers [7].

A wide range of statistical methods have been adapted for such research questions in metabolomics studies [8]. For VOC analysis, specialized review articles provide recommendations on the appropriate choice of statistical methods [7, 9]. Moreover, possible pitfalls in data analysis and adequate prevention

© 2017 IOP Publishing Ltd

strategies have also been highlighted in the literature [10, 11]. Currently, enhanced statistical methods are being used increasingly for pattern recognition in VOC studies, for instance neural networks, support vector machines and random forests [12–14]. These and similar methods are suitable for analysing patterns in VOC data sets because they consider multiple compounds of a sample simultaneously, as opposed to statistical methods like *t*-tests or Mann–Whitney *U*-tests which are applied to each compound separately. Methods which consider single compounds can help to discover biomarkers when single compounds suffice for diagnosing the disease accurately. However, considering several compounds simultaneously may reveal complex disease signatures, since relationships among compounds can be taken into account and slight changes in several compounds may add up to a pronounced distinction between groups [15].

In the present study, the presence of an induced infection was predicted from VOC data obtained in a controlled animal study. This was implemented using two opposing statistical approaches: a prediction using random forests was directly compared with a prediction based on single compounds which discriminated most clearly between healthy and infected individuals. Therefore, the actual gain from considering the whole VOC sample instead of single compounds could be evaluated for this data set. Both strategies were tested within a cross-validation procedure in order to achieve a realistic estimation of their sensitivity and specificity.

Random forests was preferred over other methods for pattern recognition since this method is non-parametric (i.e. the data are not required to conform to a given specific distribution) and robust to both outliers and correlations between compounds. In this way it also provides the possibility to include possibly correlated physiological and environmental factors (in order to control for potential confounders [10]) without overfitting the data. Moreover, in contrast to some other multivariate classification methods, random forests is also able to detect non-linear relationships between compounds and the outcome (i.e. the disease status in the present investigation) [7, 9, 16].

The animal study which provided the data for these analyses assessed differences in VOC concentrations which were related to paratuberculosis. In short, paratuberculosis, or Johne's disease, is a chronic disease in ruminants caused by infection with the bacterial pathogen *Mycobacterium avium* subsp. *paratuberculosis* (MAP). The infection results in an inflammation of the intestine and accounts for considerable economic losses in cattle farming due to reduced milk yield and slaughter value. Clinical signs like diarrhoea and severe weight loss are only apparent in the late progression of the disease after a latent phase of up to several years, whereas shedding of bacteria already starts during the subclinical phase. Hence, reliable diagnostic tests are crucial in order to single out clinically non-apparent MAP-shedding animals in a herd.

Currently the most sensitive diagnostic procedure for paratuberculosis is the cultural isolation of MAP bacteria from faecal or tissue samples. A major disadvantage of this method is the long incubation time (at least 12 weeks). Diagnosis might be accelerated by detection of MAP-specific antibodies in samples of serum or milk via enzyme-linked immunosorbent assays, especially in the case of high-shedding animals. But most importantly, all diagnostic tests available to date show a limited sensitivity, particularly in the subclinical phase. This is due to the fact that bacteria are shed intermittently and in lower amounts during this phase. In addition, faecal shedding and immune response vary individually to a large extent [17]. Therefore, there is a need for diagnostic tests with higher sensitivities and decreased processing time in order to reduce false negative findings and enable effective disease control strategies.

A recent approach for accelerating the diagnosis of paratuberculosis focuses on VOCs which are being emitted from MAP cultures with the aim of identifying *in vitro* bacterial growth earlier than currently possible by visual assessment or nucleic acid amplification techniques. By two *in vitro* experiments, MAP-related VOCs could be detected and even linked to different stages of bacterial growth [18, 19]. Nevertheless, the processing time for diagnosis might be further reduced by avoiding the limiting step of culturing and instead measuring VOCs directly *in vivo*. Pilot *in vivo* studies on goats showed that MAP-inoculated animals can be distinguished from non-inoculated animals based on VOC samples from exhaled breath or headspace over faeces by means of differential ion mobility spectroscopy (DMS) [20] and by gas chromatography-mass spectrometry (GC-MS) [21], respectively. The present animal study was based on the study design of the pilot *in vivo* study using GC-MS measurements but comprised a considerably increased sample size. From the measurements of this extensive study we aimed to assess whether VOC samples can be utilized to predict MAP infections. As the predictability of paratuberculosis by VOC samples had not been investigated so far, it was not known in advance which of the proposed data analysis strategies would be more suitable.

The focus of this publication lies in presenting two different strategies for data analysis which assist in identifying potential disease biomarkers and disease-indicative VOC profiles. Although the methodological comparison of both approaches was based on specific example data, it yielded results which may be transferred to similar studies on VOC samples. All methods are described in detail to give insight into the sequence of considerations for choosing a strategy for data analysis. Finally, we hope that this example may serve as a bridge between methodological reviews and applied data analyses.

2. Methods

2.1. Study design

Twenty-four clinically healthy goat kids were split into a case group of 14 goats and a control group of 10 goats. Animals of the case group were subsequently inoculated with MAP (strain JII-1961) by repeated administration of bacterial wet mass suspended in milk replacer. The optimum dose for inoculation had been determined experimentally in a preceding study [17].

Sampling in the course of the study covered roughly the first year of life. Faecal samples for VOC analysis were obtained from all goats approximately every 2 weeks, starting 3 weeks after completion of inoculation (wpi, weeks post-inoculation). Sampling of exhaled breath for VOC analysis was performed approximately every 4 weeks starting at 5 wpi. The last samples for VOC analysis were taken at 47 wpi (see tables S1 and S2, available online at stacks.iop.org/JBR/11/047105/mmedia). Altogether, VOC concentrations were measured for a total of 477 faecal samples and 299 breath samples. Samples of exhaled breath were partially obtained in duplicate in order to monitor the accuracy of *in vivo* measurements. In order to avoid a bias towards those repeated measurements only their averages were included. This was justified since deviations between repeated measurements were low. The resulting data set contained 238 breath samples.

Furthermore, standard diagnostics for paratuberculosis were performed in addition to VOC analyses. In order to monitor the MAP-specific interferon-γ response and seroconversion, blood samples were collected approximately every 4 weeks from all animals. Faecal shedding of MAP was examined by cultural isolation of MAP from faecal samples, which were also collected every 4 weeks. In order to examine the intestine for MAP-related pathomorphology, most animals were euthanized and dissected at the end of the experiment. According to schedule, four MAP-inoculated animals were dissected in the course of the study (at 12 and 25 wpi, respectively). As expected, MAP-positive tissues (obtained by necropsy) and faecal shedding of MAP was only observed for inoculated animals. Moreover, MAP was detected in all animals of the case group, indicating a successful inoculation. None of the inoculated animals showed clinical signs by the end of the study, thus all samples of inoculated animals represented subclinical infection.

2.2. Animals, husbandry and animal welfare

All goats were of the same breed (Thüringer Wald Ziege) and purchased at the age of 2–3 weeks from one local farm which had no reported cases of paratuberculosis in the past. After transfer to the experimental animal facility, the two groups of goats were housed in separate stables but under equal, standardized conditions. Each group contained a single female with all other goats being male. All the male goats were castrated at the beginning of the study.

Feeding was adjusted to the age of the goats. Up to the 12th week of life, goat kids were fed with commercial milk replacer. From the 9th week of life on, goats received concentrated feed (at first pellets for lambs, later on dairy concentrate pellets). Hay, water and mineral blocks were freely available over the whole course of the study. Daily clinical examinations ensured that all goats were in a good state of health.

The study was carried out in strict accordance with the German Animal Welfare Act and in conformity with the guidelines for animal welfare set forth by the European Commission. The study protocol was approved by the Committee for the Ethics of Animal Experiments and the Protection of Animals of the State of Thuringia, Germany (registration no. 04-002/12). Throughout the duration of the study, every effort was made to minimize suffering and animals were strictly treated in accordance with good veterinary practice.

2.3. Collection and measurement of samples

Samples of exhaled breath were collected, as described previously, using an automated alveolar sampling device (PAS Technology Deutschland GmbH, Magdala, Germany) combining mainstream capnometry and needle-trap microextraction (NTME; absorbent material: divinylbenzene, Carbopack X and Carboxen 1000) [22]. The device facilitates sampling of a predefined volume restricted to the alveolar phase of exhalation by considering flow rates and CO_2 thresholds [23]. It was adapted for collection of breath gas samples from goats by addition of a tightly fitting face mask [6, 21, 24]. The flow rate during sampling was 21.5 ± 1.9 ml min^{-1} (mean ± SD) and the sample volume per goat per time point was set at 50 ml.

Faecal samples were collected individually and, immediately after collection, aliquots were filled separately into 20 ml headspace vials sealed with Teflon-coated rubber septa and magnetic crimp caps (Gerstel GmbH and Co. KG, Muelheim/Ruhr, Germany). The vials were processed within 72 h of sampling. VOCs were pre-concentrated from headspace over faeces by means of solid-phase microextraction (SPME; Carboxen®/polydimethylsiloxane-SPME fibres, 75 μm, Supelco, Bellefonte, PA, USA).

Subsequently, VOCs of all gaseous samples were thermally desorbed from the needle-trap devices and SPME fibres, respectively, separated and measured using GC-MS (GC Agilent 7890A, MS Agilent 5970C inert XL MSD). Prior to measurements the GC-MS system was calibrated and optimized with adapted standard mixtures. Compounds were tentatively identified by a mass spectral library (NIST 2005, Gaithersburg, MD, USA) and selected based on concentration differences between MAP-inoculated and non-inoculated animals. In order to obtain an unequivocal identification and quantification, selected compounds were verified by measurements of pure reference substances. For the calibration and determination of the limit of detection (LOD; signal-to-noise ratio 3:1) and limit of quantification (LOQ; signal-to-noise ratio 10:1), different concentration levels of the reference substances were analysed as described previously [18, 22]. Noise was determined by repeated measurements of blank samples. Peak areas of selected compounds were calculated based on extracted ion counts using Agilent MSD Chemstation (E.02.00.493) software.

Preparation and analysis of blood samples for monitoring the immune response and of faecal samples for cultural isolation was carried out as described in a previous publication [17].

2.4. Explanation of terms and statistical procedures

The present data analysis focused on identifying if VOC samples originated from MAP-inoculated or healthy individuals. Hence, the method of choice should assign each VOC sample either to the group (or 'class') of MAP-inoculated animals or to the group of healthy animals. This is called a classification problem in the terminology of statistics. Often, classification methods are divided into supervised and unsupervised methods. These terms were first introduced in the context of machine learning. Supervised classification methods are used when the group structure is known and can be exploited to build the classifier, whereas unsupervised methods are used to discover any group structures from the data [25]. For our data analysis, supervised methods were employed.

The central point of our data analysis was to compare a classification method which considers each compound separately with a classification method which takes the complete VOC sample into consideration. In the terminology of statistics, methods using single variables are referred to as univariate whereas multivariate methods include multiple variables. The two methods that we chose for comparison are described in detail below.

As classification problems occur in diverse contexts, many different methods have been developed for different types of data. An overview on methods suited for metabolomics data, and VOC data in particular, can be found in the literature (see e.g. [7–9, 26, 27]). The choice of method should be based on the question of the study and the type of data.

For instance, some methods (referred to as parametric methods) are tailored to data that fulfil certain assumptions, for example conforming to a normal distribution. In our case, an exploratory analysis revealed that measurements on most compounds did not conform to a normal distribution and contained strong outliers. In addition, data sets contained a considerable number of zeros, especially for compounds that were only observed in concentrations close to the limit of quantification and below. For this reason, log-transforming the data to achieve a normal distribution was not an option. Instead of testing different methods of data transformation for best practice, we preferred to apply non-parametric methods, which do not require data transformation, to the original, untransformed data.

Moreover, classification methods differ in the expected type of relationship between compounds and the outcome (i.e. a linear or non-linear relationship). As we aimed to explore a potentially complex VOC pattern, we decided to choose random forests as a multivariate classification method that can also handle non-linear relationships.

In the present data analysis, classification methods are compared based on their accuracy in predicting the disease status. In order to obtain an unbiased prediction for each observation it is important to split the data into a training set (for fitting the classifier) and a test set (for evaluating the classifier), as otherwise the final estimation of sensitivity and specificity would be overoptimistic [28]. This was taken into account using five-fold cross-validation. With five-fold cross-validation, the data are first randomly split into five blocks of roughly equal size. Subsequently, in each of the five runs of the procedure, one block of data is held out (the test set) while the remaining data (the training set) are used to train the classifier. The classifier is then used for prediction on the test data set which enables us to evaluate its performance on the previously unseen data. In this way, the full data set is used in each cross-validation run such that each observation is being classified exactly once. Finally, sensitivity and specificity are estimated by averaging across all five runs.

As a general rule, the training and test set need to be independent of each other in order to achieve an unbiased estimation of sensitivity and specificity. However, the measurements of the study are not mutually independent, but actually depend on time and on the goat from which the sample originated. Since such dependences may lead to an overfitting of the cross-validation procedure [29, 30], the measurements were not assigned randomly to one of the five blocks. Instead, each goat was assigned randomly to one block so that all of the measurements on a specific goat either belonged to the training set or to the test set for each cross-validation run. In this way, we made use of the naturally occurring inter-individual variations as a test for the predictive performance of the classifiers. The group stratification was also taken into account for the assignment to ensure an equal distribution of cases and controls in the training and test set. In order to ensure that this random assignment does not obstruct the direct comparison of the accuracy of the two classification methods, the same assignment of blocks was used for both classification methods.

VOC data from exhaled breath and headspace over faeces were considered in parallel to allow a comparison of both sample sources. After defining a cross-validation scheme for each of the data sets, both classification methods were trained and evaluated analogously on both data sets.

Data analysis was performed in R version 3.4.0. with packages caTools and randomForest [31–33]. The package ggplot2 was used for visualizations [34].

2.5. Univariate classification approach

For the univariate approach, single compounds had to be selected from all available VOCs. In order to assess which of the compounds discriminated most clearly between both groups, effect sizes were calculated from the training data. In general, effect sizes measure the magnitude of variation between two groups. The advantage of effect sizes over hypothesis testing in this situation lies in their interpretability and independence of the sample size. The latter allows a comparison across different studies as opposed to *p*-values from hypothesis testing, that is why they play a key role in meta-analyses [35]. Here, the effect size was calculated as the rank-biserial coefficient of correlation based on the Mann–Whitney *U* statistic as proposed by Wendt [36]. While a large number of definitions for effect sizes exist, the advantage of this definition lies in its robustness even in the presence of strong outliers, since the non-parametric *U* statistic considers only the ranking of the values. For this definition, an effect size close to 1 corresponds to a high discrimination between both groups whereas an effect size close to 0 corresponds to a low discrimination. However, this definition of effect size does not distinguish effect directions, i.e. the group for which this compound showed higher concentrations in most cases. Therefore, this information was memorized from the previously calculated Mann–Whitney *U* statistic.

In order to gain insight into the distribution of measurements for the most promising compounds, we present combined violin and box plots for the three compounds that exhibited the highest effect sizes. Box plots summarize the spread of the data by quartiles: the interquartile range (IQR; middle 50% of the data) is depicted as a box with a horizontal line marking the median and vertical lines (‘whiskers’) extending to the most extreme value above and below the box, respectively, whose distance to the box is less than 1.5 × IQR. Measurements beyond this are depicted as single points. Violin plots visualize the distribution of measurements using a smoothed histogram which is depicted on both sides of the box plot. This adds additional information to the box plot by showing the peaks of the distribution.

After selecting the compound with the highest effect size on the respective training data set, a cut-off value was determined for this compound by means of a receiver operating characteristic (ROC) analysis based on the measurements in the training data. The cut-off value was determined such that its false positive rate (FPR; 1 − specificity) and true positive rate (TPR; sensitivity) were closest (i.e. had the shortest distance) to the optimum, namely $\mathrm{FPR} = 0$ and $\mathrm{TPR} = 1$ [37]. This cut-off value was eventually used as the decision threshold for classifying each observation of the respective test data set either to the inoculated or to the non-inoculated group.

2.6. Multivariate classification approach

Random forests was chosen as the multivariate classification method because this method does not assume a specific distribution of measurements and can handle outliers, correlations among compounds and non-linear relationships. In general, a random forest consists of a large number of decision trees, which simply use a sequence of binary (i.e. answered either ‘yes’ or ‘no’) decision rules for classification [7, 9, 16]. The decision rules can be pictured in the shape of a binary tree; an example of a decision tree is given in figure 1. The figure exemplifies a non-linear relationship between two compounds, which results in a low effect size for every single compound. Therefore, neither of the two compounds seems to qualify as a disease biomarker from this point of view. However, the decision tree gives a perfect classification without any misclassifications using just these two compounds. The decision rules are determined based on the explanatory variables (which here are also referred to as ‘features’) such that sub-regions are constructed which should each ideally contain only observations of the same class.

As decision rules may be added until perfect classification is achieved, decision trees are prone to overfitting the data. This means that a decision tree is perfectly adapted to the data it was trained on. Therefore it does not generalize well for the actual classification problem and performs badly on new data. Random forests compensates for this by restricting the construction of each decision tree on a random subsample of the training data and a random subset of all available features, which ensures a diverse set of decision trees as a result. Finally, each observation of the test data set is classified by each decision tree and is lastly assigned to the class which was reported most frequently. In this manner, it is also possible to evaluate the relative importance of each feature for the accuracy of the prediction. This is assessed during the construction of the trees by randomly resampling the measurements for each feature (i.e. replacing each measurement by a random one for this feature without consideration of the true class) using only the measurements of the training data which had not been used for constructing the tree. The resulting change in misclassification rate is averaged over all trees of the random forest [16, 33]. For strong predictors,

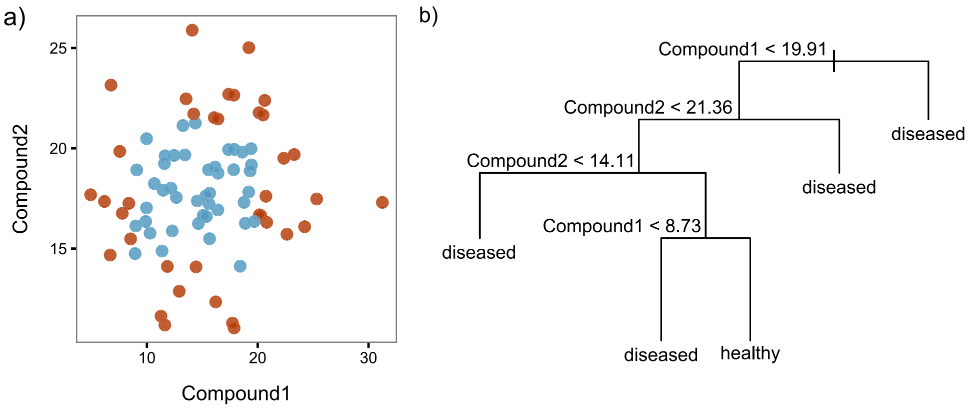

Figure 1. Artificial example of a decision tree. (a) Sample data for two artificial groups (blue healthy, red diseased) and two arbitrary compounds. None of the compounds gives an accurate prediction of the disease status when considered on its own (56.8% sensitivity and 65.1% specificity for compound 1, 59.5% sensitivity and 53.5% specificity for compound 2; values derived as described before); the effect sizes for both compounds are below 0.2. (b) However, the classification tree [38] gives a perfect classification for each observation. For example, all observations which showed concentration levels equal to or greater than 19.91 for compound 1 were correctly classified as diseased. Next, all observations which exhibited concentration levels below 19.91 for compound 1 but equal to or greater than 21.36 for compound 2 were correctly classified as diseased, etc.

resampling results in a distortion of the association with the classes. Therefore when such features are being resampled, the average misclassification rate increases considerably—in other words, the accuracy of the prediction decreases considerably. Hence, features for which resampling leads to a high mean decrease in accuracy are regarded as being of high importance for the prediction.

Apart from measurements on volatile organic compounds, additional variables that might make a contribution were included, for instance age (in weeks), body mass (in kg), rectal temperature (in °C), information on diet such as the amount (in g) and type of concentrated feed (suited for lambs or dairy goats), and for faecal samples also the volume of milk (in ml, sampling of exhaled breath started after weaning) and further variables indicating if a new mineral block was placed in the stable and if medication or treatments were necessary at the time of sampling. For example, some of the goats had to be treated for orchitis as a complication following castration. Body mass was assessed once a week, which is why the weight at the day of sampling was estimated by linear interpolation. For one goat, rectal temperature was missing for one measurement on exhaled breath and headspace over faeces, respectively. This issue was resolved by inserting the median of the last seven measurements of rectal temperature for this goat, which did not influence the results as the body temperature was in the normally observed range throughout these measurements.

For analyses by random forests, 61 features were included for exhaled breath and 56 features for headspace over faeces, respectively (all VOCs plus covariables). The number of features that were randomly selected and considered for a split decision was set to seven while constructing the decision trees. For each random forest, 500 decision trees were generated (the default value for this function). For both VOC data sets, five random forests were trained and evaluated according to the cross-validation scheme described before.

2.7. Exclusion of measurements for goat kids

Since nutrition has a major impact on the measurements, the change in feeding from milk to a plant-based diet may complicate a precise differentiation of VOC samples from MAP-inoculated and non-inoculated animals. For this reason, the data analysis was rerun as described above after excluding all VOC samples which were taken before the 20th week of life, because we assumed that the digestion of the goats was fully adapted to the plant-based diet and rumination by this age.

3. Results

Fifty-one volatile organic compounds in exhaled breath and 45 compounds in headspace over faeces were found to differ between MAP-inoculated and non-inoculated animals. These VOCs belong to the categories of aliphatic and aromatic hydrocarbons, aldehydes, esters, ketones, furans, organonitrogen and organosulfur compounds. In exhaled breath, alcohols were also detected. The majority of compounds in exhaled breath showed a decrease in concentration for MAP-inoculated goats compared with non-inoculated goats, whereas roughly two-thirds of the compounds in headspace over faeces showed (on average) higher concentrations for samples of MAP-inoculated goats than for samples of non-inoculated goats.

3.1. Prediction based on VOCs in exhaled breath

For prediction based on a single compound, 3-methylpentane was found to exhibit the highest effect size among all VOCs from exhaled breath in

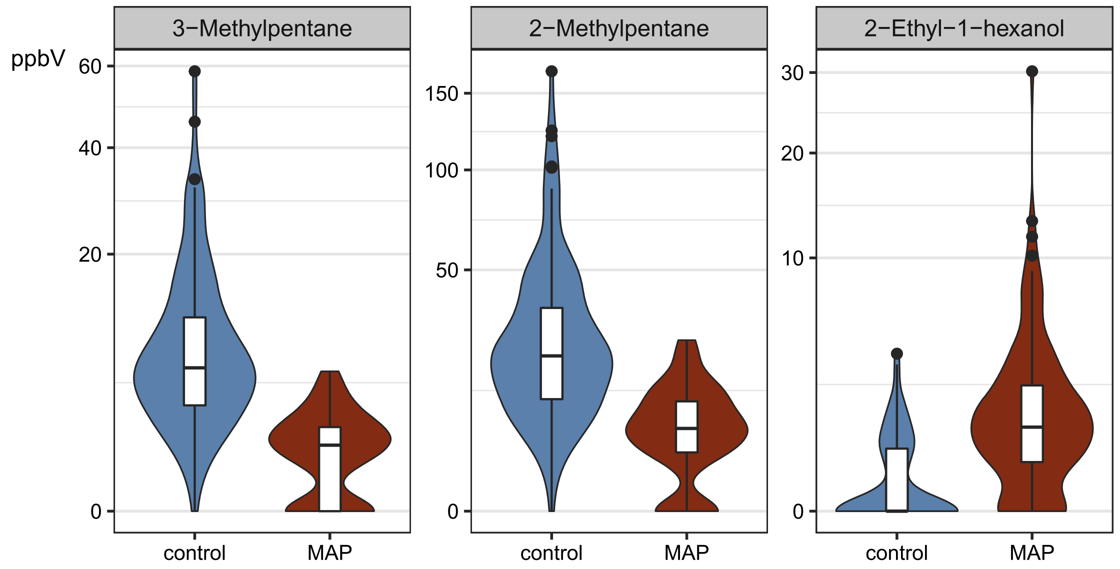

Figure 2. Comparison of the distribution of measurements for both groups for 3-methylpentane (LOD 0.13 ppbV, LOQ 0.22 ppbV), 2-methylpentane (LOD 0.93 ppbV, LOQ 1.65 ppbV) and 2-ethyl-1-hexanol (LOD 0.09 ppbV, LOQ 0.15 ppbV), respectively. These three compounds had the highest effect sizes among all volatiles that were detected in exhaled breath. The figure combines violin and box plots (box, IQR; horizontal line, median). The whiskers extend to the most extreme value above and below the box, respectively, when distance to the box is less than 1.5 × IQR. Measurements beyond this are depicted as single points. The *y*-axis is square-root transformed for better visualization.

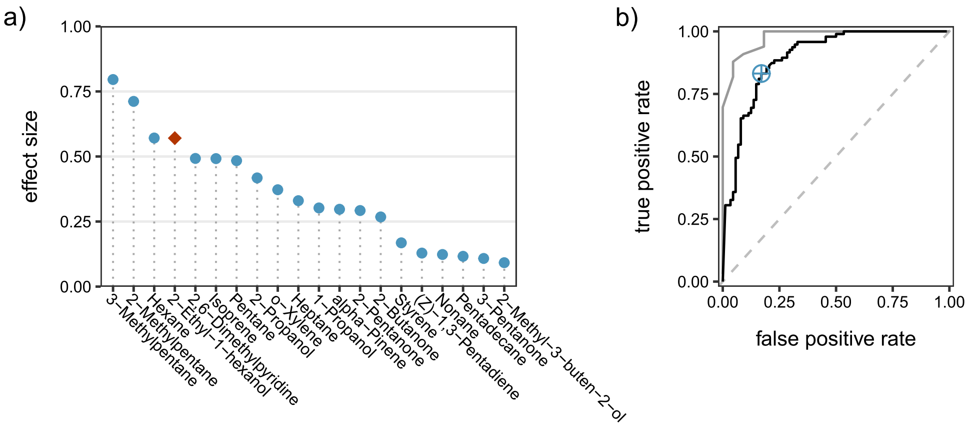

Figure 3. Results from the first cross-validation run for prediction based on single compounds from exhaled breath. (a) Effect sizes for the first 20 of the 51 detected compounds (sorted by decreasing effect size). The shape and colour of the points indicate whether high concentrations tended to be observed more frequently in MAP-inoculated goats (red rhombus) or in the control group (blue circle). (b) The ROC curve for 3-methylpentane [area under the curve (AUC) = 0.886]. The optimal cut-off (blue circle) was chosen to be 2.53 ppbV with all values above this being classified as controls and all values below being classified as MAP-positive. The solid grey line represents the respective ROC curve for the random forest from this cross-validation run.

each of the five cross-validation runs. Averaged across all five runs, 3-methylpentane reached an effect size of 0.81, followed by 2-methylpentane with an average effect size of 0.71 and 2-ethyl-1-hexanol with an average effect size of 0.57. Figure 2 depicts the distribution of the measurements for these three compounds contrasting both groups of this animal study. Obviously, the figures for 3-methylpentane and 2-methylpentane resemble each other strikingly, disregarding the differing ranges of concentration levels. 2-Ethyl-1-hexanol was one of the few compounds which tended to exhibit higher concentrations in exhaled breath of MAP-inoculated animals. In total, the rankings of the compounds according to their effect sizes were quite stable across the cross-validation procedure. Figure 3 gives an overview on the effect sizes of the top 20 compounds from exhaled breath as calculated during the first cross-validation run and depicts how the cut-off for 3-methylpentane was determined by ROC analysis during this run. As 3-methylpentane had the highest effect size in every run, prediction of the disease status was always based on this compound with varying cut-off values. The effect direction was consistent throughout the cross-validation procedure, so all measurements of the respective test set below the corresponding cut-off

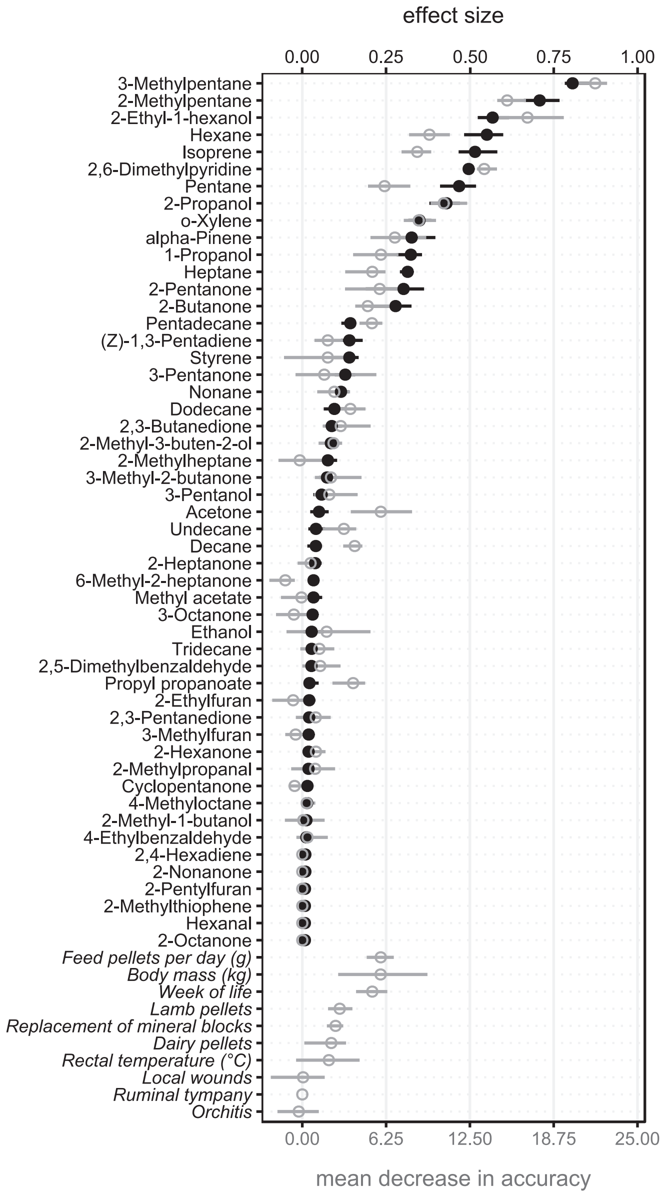

Figure 4. Contrasting estimates for exhaled breath: effect sizes for each compound (black) and mean decrease in accuracy for each feature as evaluated by random forests (grey). Circles mark the average across the five cross-validation runs, horizontal bars range from the minimum to the maximum value. Covariables which were considered only using random forests are set in italics.

value were classified as MAP-positive and all measurements greater than or equal to the cut-off were classified as MAP-negative. Averaged across all five cross-validation runs, this approach reached a sensitivity of 83.7% and a specificity of 80.0%; the confusion matrix is presented in table S3.

Consistent with the ranking according to average effect sizes, 3-methylpentane, 2-ethyl-1-hexanol and 2-methylpentane were also ranked highest regarding the respective mean decrease in accuracy caused by random resampling as estimated and averaged across the five random forests which were trained during the cross-validation procedure. Figure 4 compares the effect sizes of all compounds and the mean decrease in accuracy for all features of the random forests. All additional covariables which had been included for random forests were associated with low influences on prediction accuracy. Averaging across the whole cross-validation procedure, random forests achieved a sensitivity of 90.3% and a specificity of 81.8%. The corresponding confusion matrix is given in table S4.

3.2. Prediction based on VOCs in headspace over faeces

For VOCs in headspace over faeces, 2,3-butanedione reached the highest effect size in four out of five cross-validation runs with an average effect size of 0.50, followed by 2,3-pentanedione with an average effect size of 0.46 and isoprene with an average effect size of 0.45. Just as for VOCs in exhaled breath, the first two compounds were structurally closely related and showed a comparable distribution of measurements irrespective of the differing concentration levels (see figure 5). However, the top compounds exhibited

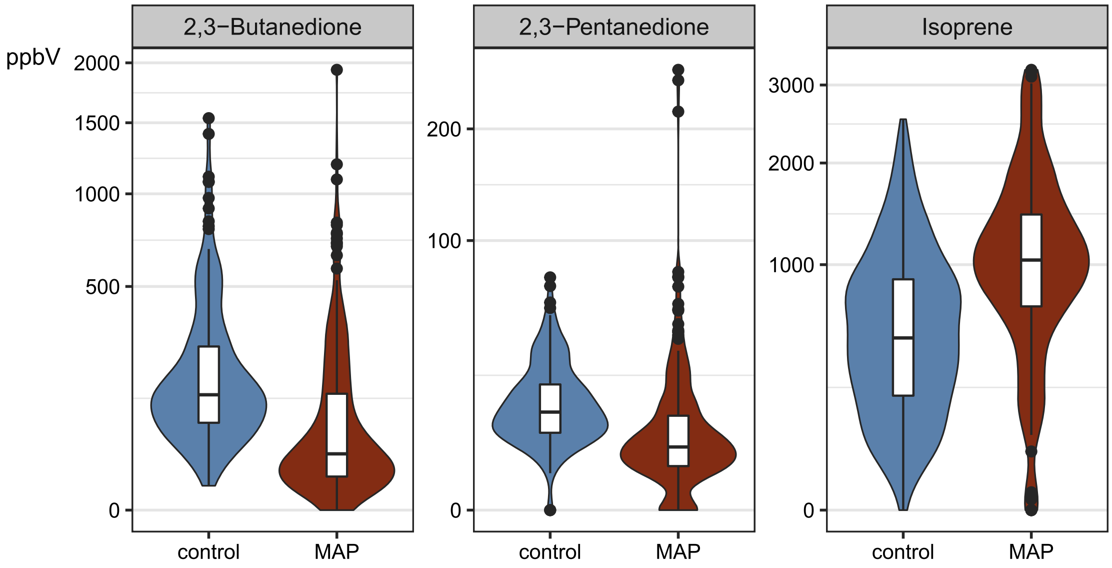

Figure 5. Comparison of the distribution of measurements for both groups for 2,3-butanedione (LOD 1.06 ppbV, LOQ 2.79 ppbV), 2,3-pentanedione (LOD 0.30 ppbV, LOQ 0.61 ppbV) and isoprene (LOD 0.26 ppbV, LOQ 0.56 ppbV), respectively. These three compounds had the highest effect sizes among all VOCs that were detected in headspace over faeces (see also figure 2).

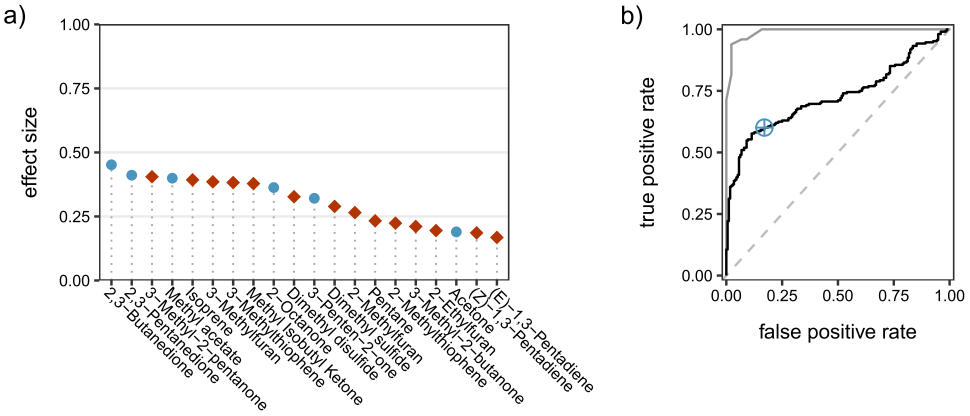

Figure 6. Results from the first cross-validation run for the prediction approach based on single compounds from headspace over faeces. (a) Effect sizes for the first 20 of all 45 detected compounds (sorted by decreasing effect size). The shape and colour of the points indicate whether high concentrations tended to be observed more frequently in MAP-inoculated goats (red rhombs) or in the control group (blue circles). (b) ROC curve for 2,3-butanedione (AUC 0.720). The optimal cut-off (blue circle) was chosen to be 59.29 ppbV with all values above being classified as controls and all values below being classified as MAP-positive. The solid grey line represents the respective ROC curve for the random forest from this cross-validation run.

considerably lower effect sizes compared with VOCs in exhaled breath. As a consequence, the rankings of the top compounds were less stable across the cross-validation procedure. That is why, finally, four-fifths of the samples were classified based on their measurements for 2,3-butanedione with varying cut-offs whereas the remaining fifth were classified based on the measurements for 3-methyl-2-pentanone. While for 2,3-butanedione all values above the respective cut-off were classified as MAP-negative, for 3-methyl-2-pentanone higher values were classified as MAP-positive, because the effect directions of these compounds were inverted. In figure 6, results for the first cross-validation run are depicted; here 2,3-butanedione was ranked first and subsequently selected for prediction. Averaged across all five cross-validation runs, this approach achieved a sensitivity of 59.7% and a specificity of 77.3%. Table S6 reports the confusion matrix.

Regarding the estimated importance of compounds as derived from random forests, 3-methylfuran was ranked as most important for the accuracy of the prediction followed by 2,3-butanedione and methyl acetate (see figure 7). Body mass was estimated to be the most important covariable among the additional features, but its influence on the decrease of prediction accuracy was again relatively low as for all covariables. Finally, a sensitivity of 86.6% and a specificity of 85.0% (averaged across the cross-validation procedure) were achieved for headspace over faeces by means

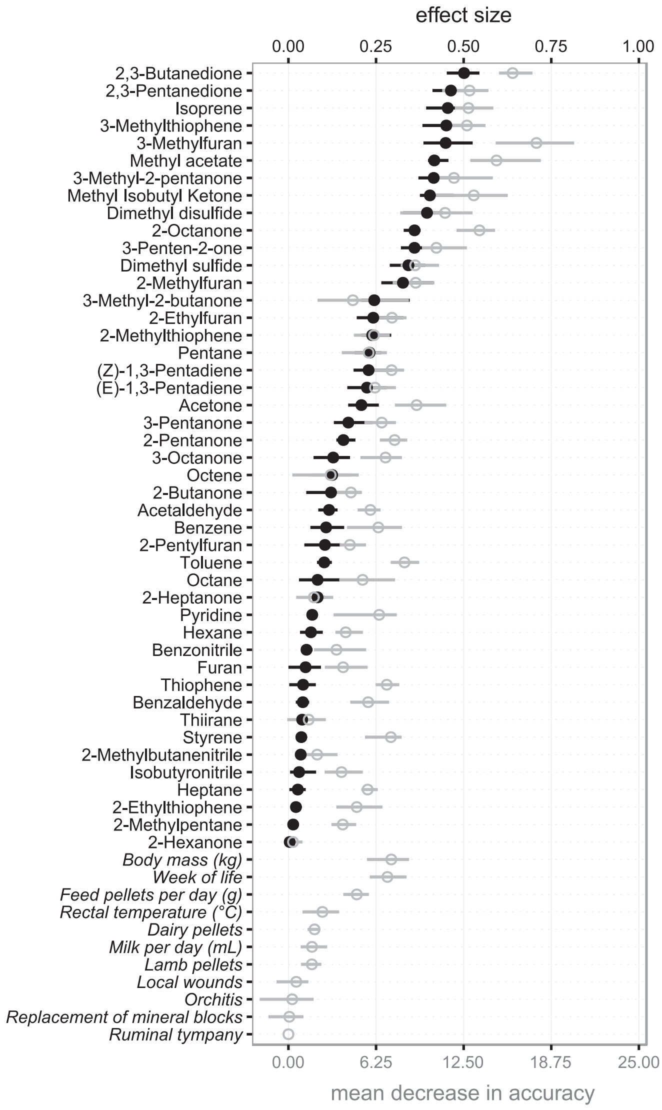

Figure 7. Effect sizes for each VOC from headspace over faeces (black) and the mean decrease in accuracy for each feature for headspace over faeces as evaluated by random forests (grey). Circles mark the average across the five cross-validation runs, horizontal bars range from the minimum to the maximum value. Covariables which were considered only using random forests are set in italics.

Table 1. Overview of sensitivity and specificity for each classification method for exhaled breath and headspace over faeces compared with the standard diagnostics for independent samples of faeces and blood.

	Exhaled breath		Headspace over faeces		Standard diagnostics		
	Single compound	Random forest	Single compound	Random forest	Faecal shedding (cultural isolation)	MAP-specific antibody response	MAP-specific IFN-γ response
Sensitivity	83.7%	90.3%	59.7%	86.6%	64.6%	48.8%	81.7%
Specificity	80.0%	81.8%	77.3%	85.0%	100%	98.6%	100%

of random forests. Table S7 reports the confusion matrix.

Overall, the predictions of the disease status by VOC measurements were less specific than the standard diagnostics which were carried out repeatedly in the course of the study (see table 1). But none of the standard diagnostics was as sensitive as the VOC based predictions by means of random forests.

3.3. Prediction after exclusion of measurements for goat kids

Exclusion of samples which were taken before the 20th week of life restricted the data set to 177 breath gas samples and 347 faecal samples. Overall, sensitivities and specificities varied only slightly compared with the original data analysis and should not be overstated (see table S9). For prediction using single compounds, sensitivities increased while specificities decreased after exclusion of these samples for both exhaled breath and headspace over faeces. For prediction using random forests, for exhaled breath sensitivity did not change at all and specificity increased, whereas for headspace over faeces sensitivity and specificity slightly decreased. Analogously, both rankings (based on effect sizes and as derived from random forests) showed also only minor changes compared with the original rankings (data not shown).

4. Discussion

Regarding VOCs in exhaled breath, the top compounds showed fairly high effect sizes. This means that these compounds showed pronounced differences in concentration between breath samples of MAP-inoculated animals and those of control animals. As a result, the prediction of the disease status based on VOCs in exhaled breath already achieved a high sensitivity and specificity using only single compounds. However, the sensitivity could be further increased by random forests. For VOCs in headspace over faeces, effect sizes of the top compounds were considerably lower than for top compounds from exhaled breath. Therefore the sensitivity of the prediction based on single compounds was also comparably low. Nevertheless, random forests increased both sensitivity and specificity for this VOC data set to such an extent that the prediction of the disease status was roughly as accurate as the predictions based on VOCs in exhaled breath.

So overall, predictions by random forests reached a higher accuracy than when considering only single compounds. It is unlikely that this increase is only due to the inclusion of potential confounding factors, since their importance for prediction accuracy was estimated to be low. The gain in accuracy using random forests points to the fact that additional information is hidden in the interrelation of the compounds which cannot be exploited when considering each compound separately [15]. Apparently, it is important to consider a VOC pattern instead of searching for single biomarkers in order to detect paratuberculosis-related signatures, particularly for measurements on headspace over faeces.

Moreover, it should be noted that the two data analysis strategies also resulted in different rankings of compounds. For instance, the ‘top five’ lists derived from both data analysis strategies for VOCs in headspace over faeces differ in three out of five compounds. For exhaled breath the top compound is 3-methylpentane for both methods, but for headspace over faeces two different VOCs are considered as top compounds (2,3-butanedione and 3-methylfuran, respectively; see tables S5 and S8). Thus it should generally be taken into consideration that candidate lists of putative biomarkers also depend on the choice of method for data analysis.

Considering the choice of statistical methods, the random forest classification method was selected as the multivariate classification method with regard to the structure of the data. However, other multivariate classification methods could have been chosen instead. The advantages of random forests in this setting have been highlighted in previous sections, but it should also be noted that while random forests can detect interrelations between features, the estimation of importance considers each feature separately and does not provide insight into the interrelations of the features [39]. Hence, it is difficult to report VOC patterns that might have been detected by random forests and may be related to the disease.

Another important consideration for multivariate classification methods revolves around the selection of variables to be included in the data analysis. We decided to include additional covariables for demonstrating how such possibly confounding factors can be assessed using random forests. As stated above, these factors contributed only slightly to the accuracy of the prediction. In fact, the low contribution of the additionally included covariables reflects the controlled conditions of the animal study as laid out in the study design. For example, feeding is known to have a high influence on VOC emissions [24], but this influence was diminished in this animal study for the reason that all goats were fed the same diet and sampling followed a standardized protocol. In addition, some of the covariables were observed only rarely (e.g. some medical treatments), so their influence on VOC emission patterns should not be extrapolated from this study. Furthermore, the controlled conditions of the animal study also contributed to the minimization of exogenous influences on VOC measurements which could not be confined for field data in the same way. For instance, environmental influences from housing and husbandry should be comparable across the whole duration of the study since animals were continuously housed in their group-specific stables under standardized conditions. On the other hand, this also means that the composition of ambient air was considerably influenced by VOC emissions from the animals. Thus indicative compounds showing increased concentrations in samples from MAP-inoculated animals might also show increased concentrations in samples of ambient air from MAP stables. Therefore we refrained from selecting only VOCs from exhaled breath which showed significant differences from ambient air. Instead, a comparison with inlet air could be made (see table S10).

Since all animals were of approximately the same age and inoculated at the same time in their first weeks of life, the course of infection was correlated with their somatic growth and metabolic changes due to the transition from goat kid to adult. For instance, the change in nutrition regime from milk replacer for goat kids towards a purely plant-based diet for adult goats entailed a major change in digestion, which was covered in this study as sampling started shortly before weaning for faeces and during weaning for breath. While this transition brings additional variation into the VOC emission patterns [6], it represents more closely the natural course of infection, because the infection is typically acquired during the first weeks of life [40]. In addition, the exclusion of VOC samples from goat kids did not result in a considerably increased sensitivity or specificity. Hence, the transition in digestion did not obscure relevant differences in VOC emissions in the present animal study.

The observation that random forests reached a higher sensitivity than the standard diagnostics which were carried out repeatedly in the course of the study raise hope that VOC measurements may in future be utilized to enhance the diagnosis of paratuberculosis. Prior to potential application, further validation by independent data is needed in order to ensure the reproducibility and stability of our results. The pilot *in vivo* study on paratuberculosis-related VOCs cannot be used directly for this aim due to differences in methodology, although the study design was comparable. This emphasizes the importance of a standardized methodology for comparing VOC measurements across different studies and settings. Nevertheless, we would like to draw some comparisons with our previous studies, which support and amend the recent findings.

In the pilot *in vivo* study on paratuberculosis-specific VOCs, 21 compounds were identified as potential biomarkers in headspace over faeces [21]. Indeed, all of them were detected in the present study as well. Although the present study considered 24 compounds in addition to those due to the increased sample size, some of the previous biomarker candidates were ranked again among the top discriminating compounds, for example 3-methylfuran, isoprene, methyl isobutyl ketone and 3-methyl-2-pentanone.

The composition of headspace over faeces is influenced in many ways by food intake, host physiology and gut microbiota, which makes it hard to trace back the origins of VOC emissions unambiguously. Nevertheless, some of the hydrocarbons, ketones and furan derivatives had been shown to be directly related to MAP by means of an *in vitro* study [19]. For instance, 3-methylfuran had been detected in significantly higher concentrations above MAP cultures compared with blank media. Thus, this compound might be detectable in higher concentrations in faeces of MAP-inoculated animals for the reason that it is emitted directly from MAP. On the other hand, 2,3-butanedione and methyl acetate were detected *in vitro* only above MAP cultures but not above blanks, indicating that these compounds might also originate from the bacteria. In our investigation both compounds were ranked among the top discriminating compounds, but the *in vivo* measurements on both compounds tended to be decreased in MAP-inoculated animals compared with non-inoculated animals. Therefore, it is plausible to assume that these compounds originate *in vivo* predominantly from other sources which are affected by the presence of MAP, such that emissions of these compounds are reduced. This gives an example why conclusions from *in vitro* measurements should not be transferred directly to *in vivo* conditions as influences from the host, its microbiome and host–microbiome interactions need to be taken into account [41].

Although the pilot study found that paratuberculosis-related differences were less pronounced in exhaled breath than in headspace over faeces [21], in the present study both classification methods achieved high sensitivities and specificities for VOC samples from exhaled breath. This coincides with the fact that the most distinctive compounds in exhaled breath from the present study had not been detected previously. Possible reasons for this may be the improved methodology or the increased sample size of the present study. The findings from exhaled breath are as yet inconsistent, but promising, so additional validation is needed which could be a starting point for further studies.

As a last point, comparing the effect sizes of the compounds that were detected in both studies reveals similarities for some of the most promising biomarker candidates from headspace over faeces such as 3-methylfuran, isoprene and 3-methyl-2-pentanone (see figure 8). Next, it is necessary to investigate whether these findings can be confirmed in farms with MAP-infected animals under naturally less controlled conditions than in this animal study. In addition, further measurements are needed to evaluate the accuracy of the VOC profiles with respect to other bacterial infections and diseases.

5. Conclusions

This article presents a data-based comparison between a univariate and a multivariate data analysis strategy in the context of VOC studies. Both strategies were used to predict whether VOC samples originated either from the experimentally infected group or from the control group. Most importantly, we were able to demonstrate that a multivariate data analysis may yield comparably precise predictions of the disease status even though initially none of the compounds seemed to qualify as a disease biomarker when considered on its own (i.e. univariate). The reason for this is the basic ability of multivariate methods to consider compounds simultaneously, which enables the detection of disease-related patterns across several compounds (i.e. disease-indicative VOC profiles). Univariate methods are not able to reveal such patterns, and are therefore ineligible for the detection of disease-indicative VOC profiles. On the other hand, results from multivariate data analyses might not be applicable to other studies, since methodological differences in study design or in VOC analysis, for example a

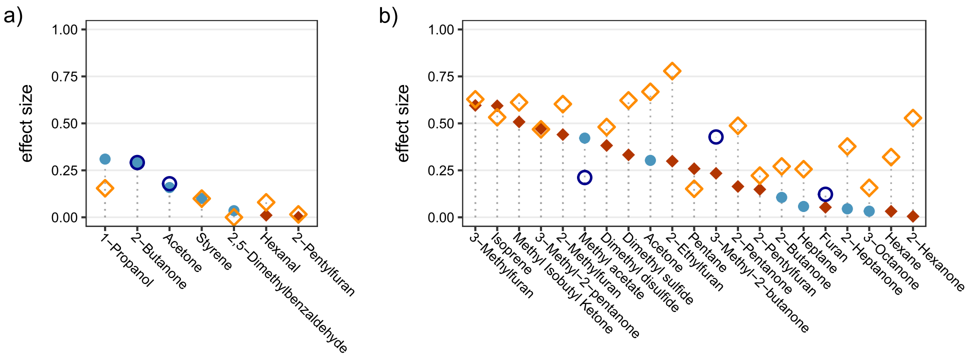

Figure 8. Effect sizes derived from the present study (solid shapes) and from the pilot *in vivo* study [21] (hollow shapes) on paratuberculosis-related VOCs for all compounds which were detected in both studies in (a) exhaled breath and (b) headspace over faeces, respectively (blue, increased in control group; red/orange, increased in MAP-inoculated animals). For comparability, measurements before the 20th week of life have been excluded for this figure.

differing composition of the adsorbent material, may entail a notable shift in concentration ranges and in covariance among compounds. In this case, similarities between studies could rather be detected using univariate statistics such as effect sizes. In addition, univariate methods are generally appropriate when single compounds are highly indicative.

In this way, both strategies provided new insights into paratuberculosis-related changes in VOC emissions. Not only could we confirm the feasibility of distinguishing MAP-inoculated goats from healthy controls by means of VOC samples, we could also exploit the distinction to predict the presence of the infection. VOC samples from both exhaled breath and headspace over faeces were well suited for prediction. As these data sets cover a long period of time, including major physiological changes in the host, it was more effective to consider the full VOC sample including potential confounders using random forests.

The proposed workflow might also be used as a template for a data analysis strategy for situations where little is known about the predictive power of single compounds and their interrelations. We argue that the random forest classification method is particularly suited for such situations for the reason that this method is able to handle non-linear relationships and correlations among compounds and to incorporate different measures on potential confounders while it is robust regarding data abnormalities and resistant to overfitting at the same time. For comparability across studies, the reporting of effect sizes is highly recommended.

Nevertheless, the high physiological variability of VOC profiles and the potential contribution of exogenous sources remain major issues that need to be addressed carefully in study design, sampling and data evaluation for VOC studies in general. In addition, we emphasize the need for a standardized methodology in order to enable a valid comparison between studies. This is essential for verifying the reproducibility of the results, and thus for defining disease biomarkers and disease-indicative VOC profiles which may eventually serve as basis for future diagnostic applications.

Acknowledgments

The authors are very grateful for the excellent assistance of laboratory and animal attendants from FLI Jena. We are also very thankful to Dietmar Hein from PAS Technology (Magdala, Germany) for his continuous technical support and to all involved members from the RoMBAT research group. Finally, we wish to extend our gratitude to the anonymous referees for the careful reading of the manuscript.

This work was funded by Deutsche Forschungsgemeinschaft (DFG, grant number RE 1098/4-2 and SCHU 1960/4-2). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of interest

None of the authors of this paper has a financial or personal relationship with other people or organizations that could inappropriately influence or bias the content of this paper.

References

[1] Pereira J, Porto-Figueira P, Cavaco C, Taunk K, Rapole S, Dhakne R, Nagarajaram H and Câmara J S 2015 Breath analysis as a potential and non-invasive frontier in disease diagnosis: an overview *Metabolites* **5** 3–55

[2] Sethi S, Nanda R and Chakraborty T 2013 Clinical application of volatile organic compound analysis for detecting infectious diseases *Clin. Microbiol. Rev.* **26** 462–75

[3] Amann A, de Lacy Costello B, Miekisch W, Schubert J K, Buszewski B, Pleil J, Ratcliffe N and Risby T 2014 The human volatilome: volatile organic compounds (VOCs) in exhaled breath, skin emanations, urine, feces and saliva *J. Breath Res.* **8** 34001

[4] van de Kant K D G, van der Sande L J T M, Jöbsis Q, van Schayck O C P and Dompeling E 2012 Clinical use of exhaled volatile organic compounds in pulmonary diseases: a systematic review *Respir. Res.* **13** 117

[5] Filipiak W *et al* 2012 Dependence of exhaled breath composition on exogenous factors, smoking habits and exposure to air pollutants *J. Breath Res.* **6** 36008

[6] Fischer S, Trefz P, Bergmann A, Steffens M, Ziller M, Miekisch W, Schubert J K, Köhler H and Reinhold P 2015 Physiological variability in volatile organic compounds (VOCs) in exhaled breath and released from faeces due to nutrition and somatic growth in a standardized caprine animal model *J. Breath Res.* **9** 27108

[7] Eckel S P, Baumbach J and Hauschild A-C 2014 On the importance of statistics in breath analysis—hope or curse? *J. Breath Res.* **8** 12001

[8] Madsen R, Lundstedt T and Trygg J 2010 Chemometrics in metabolomics—a review in human disease diagnosis *Anal. Chim. Acta* **659** 23–33

[9] Smolinska A, Hauschild A-C, Fijten R R R, Dallinga J W, Baumbach J and van Schooten F-J 2014 Current breathomics —a review on data pre-processing techniques and machine learning in metabolomics breath analysis *J. Breath Res.* **8** 27105

[10] Miekisch W, Herbig J and Schubert J K 2012 Data interpretation in breath biomarker research: pitfalls and directions *J. Breath Res.* **6** 36007

[11] Broadhurst D I and Kell D B 2006 Statistical strategies for avoiding false discoveries in metabolomics and related experiments *Metabolomics* **2** 171–96

[12] Ligor T, Pater Ł and Buszewski B 2015 Application of an artificial neural network model for selection of potential lung cancer biomarkers *J. Breath Res.* **9** 27106

[13] Aggio R B M, de Lacy Costello B, White P, Khalid T, Ratcliffe N M, Persad R and Probert C S J 2016 The use of a gas chromatography–sensor system combined with advanced statistical methods, towards the diagnosis of urological malignancies *J. Breath Res.* **10** 17106

[14] van Vliet D, Smolinska A, Jöbsis Q, Rosias P, Muris J, Dallinga J, Dompeling E and van Schooten F-J 2017 Can exhaled volatile organic compounds predict asthma exacerbations in children? *J. Breath Res.* **11** 16016

[15] Saccenti E, Hoefsloot H C J, Smilde A K, Westerhuis J A and Hendriks M M W B 2014 Reflections on univariate and multivariate analysis of metabolomics data *Metabolomics* **10** 361–74

[16] Breiman L 2001 Random forests *Mach. Learn.* **45** 5–32

[17] Köhler H, Soschinka A, Meyer M, Kather A, Reinhold P and Liebler-Tenorio E 2015 Characterization of a caprine model for the subclinical initial phase of *Mycobacterium avium* subsp. *paratuberculosis* infection *BMC Vet. Res.* **11** 74

[18] Trefz P, Koehler H, Klepik K, Moebius P, Reinhold P, Schubert J K and Miekisch W 2013 Volatile emissions from *Mycobacterium avium* subsp. *paratuberculosis* mirror bacterial growth and enable distinction of different strains *PLoS One* **8** e76868

[19] Küntzel A, Fischer S, Bergmann A, Oertel P, Steffens M, Trefz P, Miekisch W, Schubert J K, Reinhold P and Köhler H 2016 Effects of biological and methodological factors on volatile organic compound patterns during cultural growth of *Mycobacterium avium* ssp. *paratuberculosis J. Breath Res.* **10** 37103

[20] Purkhart R, Köhler H, Liebler-Tenorio E, Meyer M, Becher G, Kikowatz A and Reinhold P 2011 Chronic intestinal mycobacteria infection: discrimination via VOC analysis in exhaled breath and headspace of feces using differential ion mobility spectrometry *J. Breath Res.* **5** 27103

[21] Bergmann A *et al* 2015 In vivo volatile organic compound signatures of *Mycobacterium avium* subsp. *paratuberculosis PLoS One* **10** e0123980

[22] Trefz P, Kischkel S, Hein D, James E S, Schubert J K and Miekisch W 2012 Needle trap micro-extraction for VOC analysis: effects of packing materials and desorption parameters *J. Chromatogr.* A **1219** 29–38

[23] Trefz P, Rösner L, Hein D, Schubert J K and Miekisch W 2013 Evaluation of needle trap micro-extraction and automatic alveolar sampling for point-of-care breath analysis *Anal. Bioanal. Chem.* **405** 3105–15

[24] Fischer S, Bergmann A, Steffens M, Trefz P, Ziller M, Miekisch W, Schubert J S, Köhler H and Reinhold P 2015 Impact of food intake on *in vivo* VOC concentrations in exhaled breath assessed in a caprine animal model *J. Breath Res.* **9** 47113

[25] Hastie T, Tibshirani R and Friedman J 2009 *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (New York: Springer) (https://doi.org/10.1007/978-0-387-84858-7)

[26] Gromski P S, Xu Y, Correa E, Ellis D I, Turner M L and Goodacre R 2014 A comparative investigation of modern feature selection and classification approaches for the analysis of mass spectrometry data *Anal. Chim. Acta* **829** 1–8

[27] Xi B, Gu H, Baniasadi H and Raftery D 2014 Statistical analysis and modeling of mass spectrometry-based metabolomics data *Mass Spectrometry in Metabolomics: Methods and Protocols* ed D Raftery (New York: Springer) pp 333–53

[28] Efron B 1983 Estimating the error rate of a prediction rule: improvement on cross-validation *J. Am. Stat. Assoc.* **78** 316–31

[29] Opsomer J, Wang Y and Yang Y 2001 Nonparametric regression with correlated errors *Stat. Sci.* **16** 134–53

[30] Arlot S and Celisse A 2010 A survey of cross-validation procedures for model selection *Stat. Surv.* **4** 40–79

[31] R Core Team 2015 *R: A Language and Environment for Statistical Computing* (Vienna: R Foundation for Statistical Computing)

[32] Tuszynski J 2014 caTools: Tools: moving window statistics, GIF, Base64, ROC AUC, etc https://cran.r-project.org/web/packages/caTools/caTools.pdf

[33] Liaw A and Wiener M 2002 Classification and regression by randomForest *R News* **2** 18–22

[34] Wickham H 2009 *ggplot2: Elegant Graphics for Data Analysis* (New York: Springer) (https://doi.org/10.1007/978-0-387-98141-3)

[35] Nakagawa S and Cuthill I C 2007 Effect size, confidence interval and statistical significance: a practical guide for biologists *Biol. Rev.* **82** 591–605

[36] Wendt H W 1972 Dealing with a common problem in social science: a simplified rank-biserial coefficient of correlation based on the U statistic *Eur. J. Soc. Psychol.* **2** 463–5

[37] Perkins N J and Schisterman E F 2006 The inconsistency of ‘optimal’ cutpoints obtained using two criteria based on the receiver operating characteristic curve *Am. J. Epidemiol.* **163** 670–5

[38] Ripley B 2016 tree: Classification and regression trees https://cran.r-project.org/web/packages/tree/tree.pdf

[39] Wright M N, Ziegler A and König I R 2016 Do little interactions get lost in dark random forests ? *BMC Bioinform.* **17** 145

[40] Sweeney R W 2011 Pathogenesis of paratuberculosis *Vet. Clin. North Am. Food Anim. Pract.* **27** 537–46

[41] Zhu J, Bean H D, Wargo M J, Leclair L W and Hill J E 2013 Detecting bacterial lung infections: *in vivo* evaluation of *in vitro* volatile fingerprints *J. Breath Res.* **7** 16003