Preliminary investigation of human exhaled breath for tuberculosis diagnosis by multidimensional gas chromatography – Time of flight mass spectrometry and machine learning

Marco Beccaria, Theodore R. Mellors, Jacky S. Petion, Christiaan A. Rees, Mavra Nasir, Hannah K. Systrom, Jean W. Sairistil, Marc-Antoine Jean-Juste, Vanessa Rivera, Kerline Lavoile, Patrice Severe, Jean W. Pape, Peter F. Wright, Jane E. Hill

PII: S1570-0232(17)31882-2
DOI: https://doi.org/10.1016/j.jchromb.2018.01.004
Reference: CHROMB 20985

To appear in:

Received date: 3 November 2017
Revised date: 31 December 2017
Accepted date: 3 January 2018

Please cite this article as: Marco Beccaria, Theodore R. Mellors, Jacky S. Petion, Christiaan A. Rees, Mavra Nasir, Hannah K. Systrom, Jean W. Sairistil, Marc-Antoine Jean-Juste, Vanessa Rivera, Kerline Lavoile, Patrice Severe, Jean W. Pape, Peter F. Wright, Jane E. Hill , Preliminary investigation of human exhaled breath for tuberculosis diagnosis by multidimensional gas chromatography – Time of flight mass spectrometry and machine learning. The address for the corresponding author was captured as affiliation for all authors. Please check if appropriate. Chromb(2017), https://doi.org/10.1016/j.jchromb.2018.01.004

Preliminary Investigation of Human Exhaled Breath for Tuberculosis Diagnosis by Multidimensional Gas Chromatography – Time of Flight Mass Spectrometry and Machine Learning

Marco Beccaria[1], Theodore R. Mellors[1], Jacky S. Petion[2,3], Christiaan A. Rees[4], Mavra Nasir[4], Hannah K. Systrom[4], Jean W. Sairistil[2,3], Marc-Antoine Jean-Juste[2,3], Vanessa Rivera[2,3], Kerline Lavoile[2,3], Patrice Severe[2,3], Jean W. Pape[2,3], Peter F. Wright[5], Jane E. Hill[1,4*]

[1]Thayer School of Engineering, Dartmouth College Hanover NH, USA

[2]Groupe Haitien d'Etude du Sarcome de Kaposi et des Infections Opportunistes (GHESKIO), Port-au-Prince, Haiti

[3]Department of Medicine of Weill Cornell Medical College, New York, NY

[4]Geisel School of Medicine, Dartmouth College, Hanover NH, USA

[5]Division of Infectious Disease and International Health, Dartmouth Hitchcock Medical Center, Lebanon NH, USA

*Corresponding author: jane.e.hill@dartmouth.edu

ABSTRACT

Tuberculosis (TB) remains a global public health malady that claims almost 1.8 million lives annually. Diagnosis of TB represents perhaps one of the most challenging aspects of Tuberculosis control. Gold standards for diagnosis of active TB (culture and nucleic acid amplification) are sputum-dependent, however, in up to a third of TB cases, an adequate biological sputum sample is not readily available. The analysis of exhaled breath, as an alternative to sputum-dependent tests, has the potential to provide a simple, fast, and non-invasive, and ready-available diagnostic service that could positively change TB detection. Human breath has been evaluated in the setting of active tuberculosis using thermal desorption-comprehensive two-dimensional gas chromatography–time of flight mass spectrometry methodology. From the entire spectrum of volatile metabolites in breath, three random forest machine learning models were applied leading to the generation of a panel of 46 breath features. The twenty-two common features within each random forest model used were selected as a set that could distinguish subjects with confirmed pulmonary *M. tuberculosis* infection and people with other pathologies than TB.

Keywords

Breath analysis; comprehensive two-dimensional gas chromatography; pulmonary tuberculosis; machine learning; volatile organic compounds.

1. INTRODUCTION

According to the World Health Organization (WHO), Tuberculosis (TB) is among the top ten leading causes of death worldwide [1]. TB is caused by *Mycobacterium tuberculosis* (Mtb), a bacteria that most commonly infects the lungs. Active TB disease results in generic symptoms such as cough, fever, night sweats, or weight loss, which are not specific and may be mild for many months prior to clinical evaluation. This delay can lead to dissemination of the disease within the population, as well as an increased risk of morbidity and mortality for the affected individual. The gold standard for TB diagnosis requires the culturing of the bacterium from patient sputum, where more than 80% of all TB cases are identified with a specificity of almost 100% [1,2]. In many countries, Mtb DNA in sputum is amplified (*e.g.*, GeneXpert MTB/RIF®), as a first step to guide treatment decisions within three days of the patient's clinical presentation, a process that has been recommended by the WHO since 2010 [1]. The collection of expectorated sputum samples is considered to be relatively non-invasive, however, for patients who cannot easily produce sputum (such as children or individuals co-infected with human immunodeficiency virus (HIV)), more invasive sampling approaches, including induced sputum or gastric aspirate, are required. Therefore alternative samples that can be collected non-invasively, such as urine [3] and exhaled breath [4], present an attractive target for future TB diagnostics [5].

The evidence that diagnostic volatile molecules are produced during Mtb infection come from studies using rats that correctly sniffed and classified over forty thousand sputum samples originating from over ten thousand individuals [6]. The biomarker(s) that the rats are detecting remain unknown, although the signal is almost certainly olfactory, suggesting that the molecules exist in the gas phase at ambient temperature. Several studies have investigated the volatile molecules produced in breath during lung infection with *M. tuberculosis* using gas-chromatography mass-spectrometry (GC-MS) [7-11]. Each of these studies report different sets of marker compounds, possibly due to patient heterogeneity (including co-morbidities) as well as

different stages of pathogenesis. Indeed, volatile organic compounds (VOCs) possibly produced by Mtb may be released or modified by the host at different times [12] and they may be present at low concentration [11].

Comprehensive two-dimensional gas chromatography (GC×GC) is a good candidate for enhanced biomarker discovery due to the increase in peak capacity and sensitivity (when a cryogenic modulator is used) compared to a monodimensional GC. The GC×GC system employs two orthogonal mechanisms to separate the constituents of the sample within a single analysis by using two columns coated with different stationary phases. Sensitivity and limits of detection are improved due to focusing of the peak in the modulator and separation of analytes from chemical background compared 1D-GC [13]. In 2006, Libardoni used for the first time a GC×GC system coupled to flame ionization detector (FID) to human breath collected in a multi-bed thermal desorption (TD) trap [14] and in the same year, Sanchez developed a similar TD-GC×GC method coupled to Time-of-flight mass spectrometer (ToFMS) [15]. Since then, GC×GC methodology was used to analyze human breath in several studies [16-18]. The number of volatile molecules detected by GC×GC-MS for breath analysis is generally 8-10-fold higher compared to mono-dimensional GC-MS [16-17]. Moreover, ToFMS technology has several advantages, such as, full mass spectra acquisition at trace level sensitivity and mass spectral continuity, which allows for deconvolution of spectra of co-eluted peaks [19].

The aim of this pilot study was to assess the capability of GC×GC-(ToF)MS methodology for discriminating between subjects with confirmed pulmonary Mtb infection and people no Mtb infected using their breath. To the best of our knowledge, GC×GC-(ToF)MS has never previously been employed to study pulmonary TB via the analysis of exhaled breath in humans.

2. EXPERIMENTAL

2.1 Patient demographics and tuberculosis infection confirmation

A total of 34 individuals, including 14 with active pulmonary TB and 20 with sexually transmitted infections (STI) (female in reproductive age with no-TB symptom/infection) as control group, presenting to GHESKIO Clinical Center (Port-au-Prince, Haiti) between September and October of 2016 were included in the present study. Sputum samples were collected following WHO

guidelines [20]. An Institutional Review Board at the collaborating site approved the research. All subjects gave their signed informed consent to participate and were at least 18 years old. All patients classified as TB-positive lung disease had both a positive acid-fast bacillus testing and GeneXpert MTB/RIF® result. Patient demographic information is reported in Table 1.

<insert Table 1>

2.2 Materials and analytical instrumentation

One liter Tedlar bags (SKC Inc, Eighy Four, PA, US) were used for the collection of breath [22].

A three-bed TD tube containing Carbopack Y, X, and Carboxen 1000 (Supelco, Bellefonte, PA), a sorbent combination previously optimized for the collection of a wide range of breath molecules, was used to concentrate and store volatile molecules [15]. TD tubes were desorbed in a Pegasus 4D (LECO Corporation, St. Joseph, MI) GC×GC-ToFMS instrument with an Agilent 7890 GC equipped with a TDU, CIS, and MPS autosampler (Gerstel, Linthicum Heights, MD). Solvent venting time: 10 min (30°C; 60 mL/min); cryofocusing time: 5 min (-100°C), sample desorption time: 180s; CIS temperature: 330°C; injection mode: splitless. Chromatographic analysis was performed using a Rxi-624Sil (60m×250μm×1.4μm) as ^{1}D-GC column and a Stabilwax (1.5m×250μm×0.5μm) as ^{2}D-GC column, both purchased from Restek (Bellafonte, PA, US). Modulation time was 2 sec total and helium as carrier gas (flow-rate: 2 mL/min). For detecting breath molecules, a ToFMS was employed, with the following parameters: electron impact at 70 eV; acquisition range: 30-500 *m/z*; acquisition rate: 200 spectra/s; ion source temperature: 200°C.

2.3 Breath and room air sampling

Briefly, before breath collection, patients and volunteers rinsed their mouth with water and exhaled normally for 1-2 sec into the room. Before breath collection, all bags were thoroughly flushed with high purity nitrogen gas to remove residual contaminants [21]. Subjects exhaled normally into a 1L-Tedlar bag via a mouthpiece and exhaled breath was sorbed onto a TD tube hermetically sealed

and stored until further analysis, as previously reported [22]. Room air was also collected in the same manner.

2.4 Chromatographic alignment and feature identification

Baseline correction and peak picking of all acquired data files were carried out using ChromaTOF software. All processed data files (n = 44: 34 breath samples + 10 room air samples) were first aligned on the basis of retention times and mass spectral matching using statistical compare feature of ChromaTOF. Tentative identification of the molecules was carried out by NIST library search (NIST library 2011). Peaks with at least signal-to-noise (S/N) ≥ 50 (S/N ≥ 20 for sub peak) and a forward match score ≥ 850/1000 for putative peak identification were required and a score ≥ 700/1000 to match the peaks. The maximum retention time difference was set at 2 s in the first dimension and 0.2 s in the second dimension.

2.5 Multivariate analysis

All statistical analyses were performed with R software version 3.3.2. Peaks eluting before 350 s were removed from the peak table. Prior to statistical analysis, breath and room air samples were normalized by Probabilistic Quotient Normalization [23]. Peak intensities were (1) log-transformed, (2) mean-centered, and (3) unit-scaled. For the identification of discriminatory features between TB-positive and TB-negative samples, the random forest (RF) classification algorithm was used [24]. This machine learning method constructs numerous de-correlated decision trees to classify samples. Features were selected as discriminatory if they were ranked in the top 100 in 80 % of RF iterations of 1,000 trees, based on their mean decrease in impurity. Principal components analysis (PCA) [25] was used to visualize the variance between samples in the data set. Feature significance was tested using the Mann- Whitney *U*-test [26], with Benjamini-Hochberg correction [27]. A significance level of $p < 0.05$ was selected.

3. RESULTS AND DISCUSSION

3.1 Composition of exhaled breath

Across the 34 breath samples analyzed, a total of 2,549 features were detected. Contaminants and artifacts (e.g., siloxanes, phthalates) were removed when identified with a spectral similarity ≥ 850/1,000 relative to the NIST 2011 library [28]. Each feature with a similarity score < 850 was manually checked and deleted when associated with a contaminant or artifact, resulting in a reduction to 1,513 features. The matrix obtained was further reduced using different frequency of observation (FOO) cutoffs where: features present in at least 20 % (n = 1228), 50 % (n = 792), and 80 % (n = 515) of samples within one class were retained for statistical evaluation. A two-class (TB-positive versus control) RF analysis [24] was applied for each set of FOOs. Features were defined as discriminatory if they ranked in the top 80/100 features (as defined by their mean decrease in impurity) in every RF iteration. Twenty-nine, 35 and 39 features were selected as discriminatory through RF classification, for 20%, 50%, and 80% FOO cutoffs respectively, obtaining a panel of 46 features. Twenty-two out of these 46 features were found to be in common for each model and selected for further statistical evaluation. Figure 1 reports (a) a step-by-step schema for feature reduction, (b) Venn diagram of the distribution of 46 features in three RF models with 22 features in common, and (c) chemical class of a panel of 46 features.

<insert Figure 1>

3.2 Breath evaluation and selected molecules

Across the breath samples, a wide range of chemical classes were identified, with hydrocarbons being the most commonly-identified chemical class in human breath, as reported previously [7-11,14-18,29]. The panel of 46 features obtained after different FOO and 80/100 RF cutoffs (Figure 1a), consisted mostly of hydrocarbons (52%), followed by aromatics (11%), halogen-containing compounds (9%), sulfur-containing compounds (7%), ketones (5%), and nitrogen-containing compounds (4%). Alcohols, aldehydes, esters, and unknowns represented ≤ 2% each.

Of the 46 discriminatory features identified in the three RF models, 22 human breath features were in common to all three FOO cutoffs, and eight of these were assigned a putative identity. Bubble plot of the 22 selected features as discriminatory, representing the two-dimensional GC separation

in human breath according to their retention times in ^{1}D- and ^{2}D-GC columns, is reported in Figure 2.

<insert Figure 2>

One of these compounds was previously reported as a TB-associated human breath biomarker: 2-butyl-1-octanol [8], while five (benzene, 3-methyl-thiophene, 2-4-dimethyl-hexane, 2,3-dimethyl-heptane, methacrolein), have been previously reported by GC(×GC)-MS in humans as breath molecules [18,29]. Methanesulphonylchloride has been detected in healthy humans as a volatile molecule produced by skin, but never in human breath [29]. To the best of our knowledge, 4-ethyl-5-methyl-nonane has not been previously reported as human breath molecule, although high resolution MS and/or the utilization of pure standards would be needed for the confirmation of molecule's identity.

The ability of volatile metabolites to discriminate between TB-positive and TB-negative subjects was assessed using RF. A two-class supervised RF was used to select volatile biomarkers and then an unsupervised RF proximity model (TB-positive versus TB-negative versus room air) was created on the discriminatory volatiles selected from the two-class model [30]. The proximities from the unsupervised RF were used to plot a PC score plot (Figure 3).

<insert Figure 3>

In the PC scores plot generated from unsupervised RF proximities, room air samples are shown to cluster together, while TB-positive samples showed the highest variance, and one sample is spread near to the TB-negative samples (Figure 3). TB negative subjects are tightly clustered in the PC space.

Figure 4 reports the expression of the 22 discriminatory breath molecules in TB positive and negative classes. The bar color represents $\log_{10}$ of the difference of the normalized area of the

average of each discriminatory molecule used between TB positive and negative population. For each feature, a putative compound identification or chemical class assignment, when the identification was not possible, is reported. Feature 20 (2-buthy-1-Octanol) was the most expressed in the breath of TB positive population and was detected in 85% of this group (12/14), while only in 50% in control group (10/20). This breath molecule was previous reported by Phillips et al [8] as possible biomarker in human Mtb infection. The hydrocarbons were more abundant in the control group, except for feature 21. Feature 15 (alcohol) was most abundant in TB positive group, although it was observed in 57% (8/14) of this group against in 100% of the control group (20/20). Feature 22 (halogen containing molecule) was more often observed in the exhaled breath of control group (19/20) relative to the TB-positive study group (9/14). The other features were observed in least at 95% of samples in both populations.

<insert Figure 4>

4. CONCLUSION

This work is a preliminary step of a larger project evaluating the use of breath for the identification of TB lung pathologies. To the best of the authors' knowledge, for the first time human breath has been evaluated in the setting of TB using the GC × GC(-ToF)MS methodology, showing the capability to distinguish between pulmonary Mtb infection and people with other pathologies different than TB. From the entire spectrum of volatile metabolites in breath, three RF models were applied, leading a panel of 46 breath features. The twenty-two common features within each RF model used were selected as a set that could distinguish between Mtb-infected and -uninfected population. Eight out of 22 were tentatively identified and one of them has been resulting in common with the possible biomarkers reported in previous studies on human breath with TB infection. Although the discrimination obtained by multivariate elaboration seems very promising to extrapolate diagnostic information, minimizing the risk of experimental error in the primary identification of breath biomarkers, requires studies with a large number of individuals with TB-suspects as a control group (people with TB symptoms but not Mtb-infected) including subjects with confounding factors (as HIV co-infection).

Acknowledgements

Christiaan A. Rees was supported by the National Institutes of Health training grant (Grant#T32LM012204). Mavra Nasir was supported by the Burroughs Welcome Fund institutional program grant unifying population and laboratory based sciences to Dartmouth College (Grant#1014106).

References

[1] World Health Organization (WHO), Global Tuberculosis Report 2016, pp. 1-214.

[2] J.J. Lee, J. Suo, C.B. Lin, J.D. Wang, T.Y. Lin, Y.C. Tsai, Comparative evaluation of the BACTEC MGIT 960 system with solid medium for isolation of mycobacteria, Intern. J. Tubercolosis Lung Disease 7 (2003) 5698-574.

[3] S.D. Lawn, Point-of-care detection of lipoarabinomannan (LAM) in urine for diagnosis of HIV-associated tuberculosis: a state of the art review, BMC Infect. Diseases, 12 (2012) 103

[4] A.W. Boots, J.J.B.N. van Berkel, J.W. Dallinga, A. Smolinska, E.F. Wouters, F.J. van Schooten, The versatile use of exhaled volatile organic compounds in human health and disease, J. Breath Res. 6 (2012) 1-21.

[5] A. Amann, W. Miekish, J. Shubert, B. Buszewski, T. Ligor, T. Jezierskik, J. Pleil, T. Risby, Analysis of exhaled breath for disease detection. Annu. Rev. Anal. Chem. 7 (2014) 455-482.

[6] A. Poling, A. Mahoney, N. Beyene, G. Mgode, B. Weetjens, C. Cox, A. Durgin, Using giant african pouched rats to detect human tuberculosis: a review, Pan. Afr. Med J. 21 (2015) 333.=

[7] M. Phillips, R.N. Cataneo, R. Condos, G.A. Ring Erickson, J. Greenberg, V. La Bombardi, M.I. Munawar, O. Tietje, Volatile biomarkers of pulmonary tuberculosis in the breath, Tuberculosis 87 (2007) 44–52.

[8] M. Phillips, V. Basa-Dalay, G. Bothamley, R.N. Cataneo, P.K. Lam, M.P.R. Natividad, P. Schmitt, J. Wai, Breath biomarkers of active pulmonary tuberculosis, Tuberculosis 90 (2010) 145–51.

[9] M. Phillips, V. Basa-Dalay, J. Blais, G. Bothamley, A. Chaturvedi, K.D. Modi, M. Pandya, M.P.R. Natividad, U. Patel, N.N. Ramraje, P. Schmitt, Z.F. Udwadia, Point-of-care breath test for biomarkers of active pulmonary tuberculosis, Tuberculosis 92 (2012) 314–20.

[10] A.H. Kolk, J.J.B.N. van Berkel, M.M. Claassens, E. Walters, S. Kuijper, J.W. Dallinga, F. van Schooten, Breath analysis as a potential diagnostic tool for tuberculosis, Int. J. Tuberc. Lung Dis. 16 (2012) 777-782.

[11] N.A. Dang, H.G. Janssen, Arend H.J. Kolk, Rapid diagnosis of TB using GC–MS and chemometrics, Bioanalysis 5 (2013) 3079-3097.

[12] H.D. Bean, J. Jiménez-Díaz, J. Zhu, J.E. Hill, Breathprints of model murine bacterial lung infections are linked with immune response, Eur. Resp. J. 45 (2015) 181-190

[13] P.Q. Tranchida, G. Purcaro, P. Dugo, L. Mondello, The modulators for comprehensive two-dimensional gas chromatography: a review, Trends Anal Chem. 30 (2011) 1437-1461.

[14] M. Libardoni, P.T. Stevens, J.H. Waite, R. Sacks, Analysis of human breath samples with a multi-bed sorption trap and comprehensive two-dimensional gas chromatography (GC×GC), J Chromatogr. B 842 (2006) 13–21.

[15] J.M. Sanchez, R.D. Sacks, Development of a Multibed Sorption Trap, Chomprehensive Two-Dimensional Gas Chromatography, and Time-of-Flight Mass Spectrometry System for the Analysis of Volatile Organic Compounds in Human Breath, Anal. Chem, 78 (2006) 3046-3054

[16] M. Caldeira, R. Perestrelo, A.S. Barros, M.J. Bilelo, A. Morête, J.S. Câmara, S.M. Rocha, Allergic asthma exhaled human breath metabolome: A challenge for comprehensive 2D-GC-TOF, J. Chromatogr. A 1254 (2012) 87-97.

[17] M. Phillips , R.N. Cataneo, A. Chaturvedi, P.D. Kaplan, M. Libardoni, M. Mundada, U. Patel, X. Zhang, Detection of an Extended Human Volatome with Comprehensive Two-Dimensional Gas Chromatography Time-of-Flight Mass Spectrometry, PLOSE ONE, 8 (2013) 1-8.

[18] M.K. Das, S.C. Bishwal, A. Das, D. Dabral, A. Varshney, V.K. Badireddy, R. Nanda, Investigation of Gender-Specific Exhaled Breath Volatome in Humans by GCxGC-TOFMS, Anal. Chem. 86 (2014) 1229-1237.

[19] R. Shellie, P.J. Marriot, P. Morrison, Concepts and preliminary observations on the triple-dimensional analysis of complex volatile samples by using GC×GC−TOFMS, Anal. Chem. 73 (2001) 1336-1344.

[20] World Health Organization (WHO). Guidelines for the Prevention of Tuberculosis in Health Care Facilities in Resource-Limited Settings. Geneva, World Health Organization, 1999.

[21] P. Mochalski, B. Wzorek, I. Sliwka, A. Amann, Suitability of Different Polymer Bags for Storage of Volatile Sulphur Compounds Relevant to Breath Analysis, J. Chromatogr. B 877 (2009) 189-196.

[22] T.R. Mellors, L. Blanchet, J.L. Flynn, J. Tomko, M. O'Malley, C.A. Scanga, P.L. Lin, J. E. Hill, 2017 A new method to evaluate macaque health using exhaled breath: A case study of M. tuberculosis in a BSL-3 setting, J. Appl. Physiol. 122 (2017) 695-701.

[23] F. Dieterle, A. Ross, G. Schlotterbeck, H. Senn, Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabolomics, Anal. Chem. 78 (2006) 4281–4290.

[24] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5-32.

[25] T.J. Hastie, R.J. Tibshirani, J. Friedman, The Elements of Statistical Learning. Data Mining, Inference and Prediction. New York: Springer, 2001.

[26] H.B. Mann, D.R. Whitney, On a test of whether one of two random variables is stochastically larger than the other, Ann. Math. Stat. 18 (1947) 50–60.

[27] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. R. Stat. Soc. B 57 (1995) 289–300.

[28] H.D. Bea n, J-M.D Dimandja, J.E. Hill, Bacterial volatile discovery using solid phase microextraction and comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry, J. Chromatogr. B 901 (2012) 41–46.

[29] B.D.L. Costello, A. Amann, H. Al-Kateb, C. Flynn, W. Filipiak, T. Khalid, D. Osborne, N.M. Ratcliffe, A review of the volatiles from the healthy, J. Breath Res. 8 (2014) 014001

[30] A. Smolinska, E.M.M. Klaassen, J.W. Dalinga, K.D.G. van de Kant, Q. Jobsis, E.J.C. Moonen, O.C.P. van Shayck, E. Dompeling, F.J. van Schooten, Profiling of Volatile Organic Compounds in Exhaled Breath As a Strategy to Find Early Predictive Signatures of Asthma in Children, PLOS ONE 9 (2014) e95668.

Table 1. Characteristics of the study population

	M. tuberculosis (+)	***M. tuberculosis (-)***
	(n = 34)	
Number (%)	14 (47%)	20 (53%)
Age, mean (±SD)	32 (±14)	31 (±10)
Gender (M/F)	3/11	0/20
TB Symptoms (Y/N)	13/1	0/20

Highlights

- For the first time a multidimensional comprehensive gas chromatography system was used to evaluate human breath in the setting of tuberculosis disease
- Three random forest models were applied for data elaboration
- Twenty-two common features within each random forest model used were selected as a set to distinguish between infected and control group

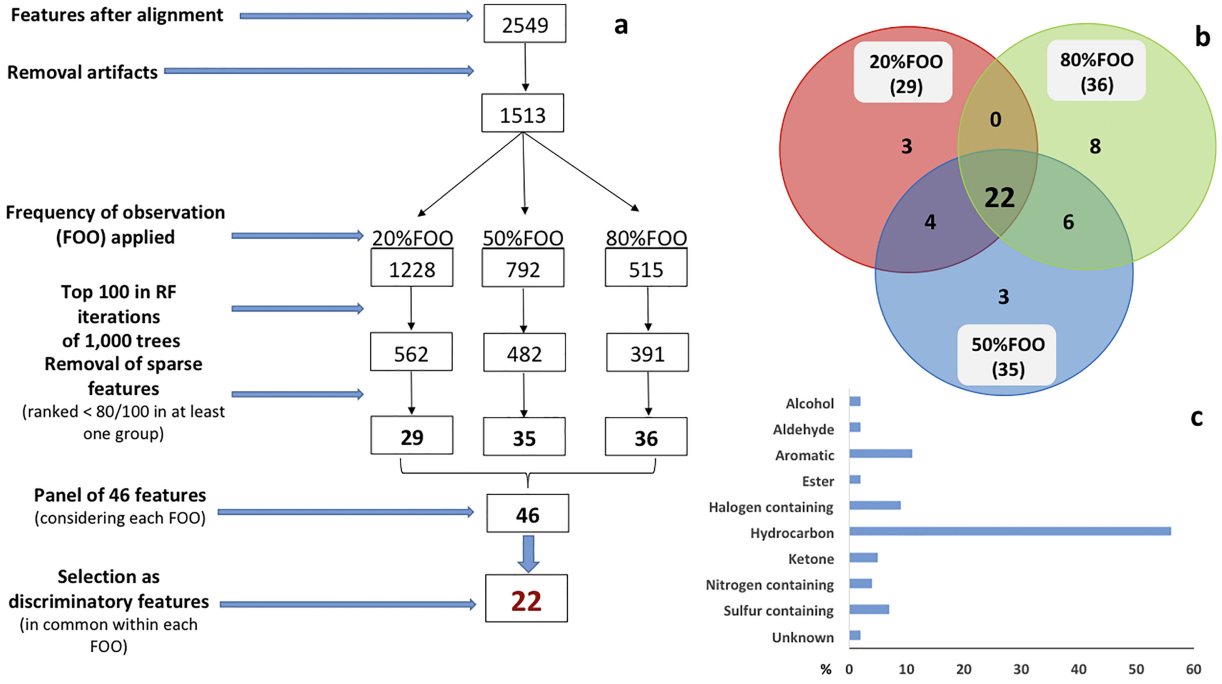
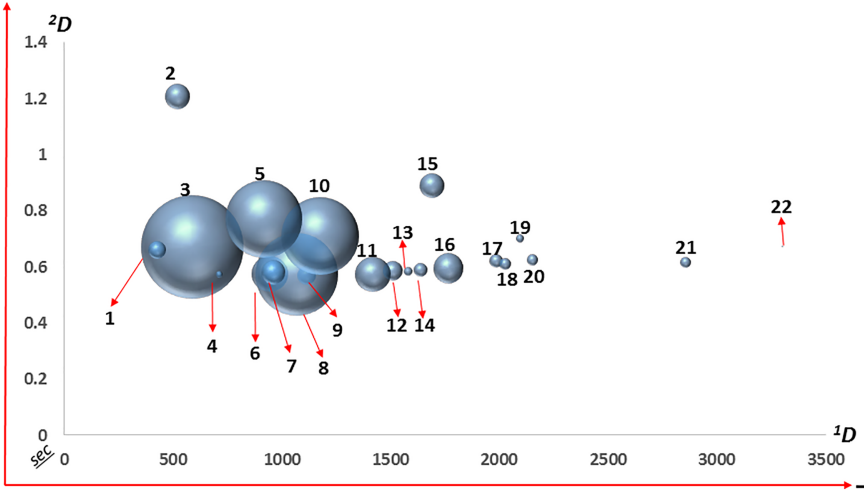

a
Features after alignment
2549
Removal artifacts
1513
Frequency of observation (FOO) applied
20%FOO
50%FOO
80%FOO
1228
792
515
Top 100 in RF iterations of 1,000 trees
Removal of sparse features
(ranked < 80/100 in at least one group)
562
482
391
29
35
36
Panel of 46 features
(considering each FOO)
46
Selection as discriminatory features
(in common within each FOO)
22
b
20%FOO (29)
80%FOO (36)
50%FOO (35)
3
0
8
22
4
6
3
c
Alcohol
Aldehyde
Aromatic
Ester
Halogen containing
Hydrocarbon
Ketone
Nitrogen containing
Sulfur containing
Unknown
%
0
10
20
30
40
50
60

Figure 1



1	Methacrolein
2	Methanesulfonyl-chloride
3	Benzene
4	2,4-dimethyl-hexane
5	3-methyl-thiophene
6	2,3-dimethyl-heptane
7	Hydrocarbon
8	4-ethyl-5-methyl-nonane
9	Hydrocarbon
10	Aromatic
11	Hydrocarbon
12	Sulfur containing
13	Hydrocarbon
14	Hydrocarbon
15	Alcohol
16	Hydrocarbon
17	Hydrocarbon
18	Hydrocarbon
19	Hydrocarbon
20	2-butyl-1-octanol
21	Hydrocarbon
22	Halogen containing

Figure 2

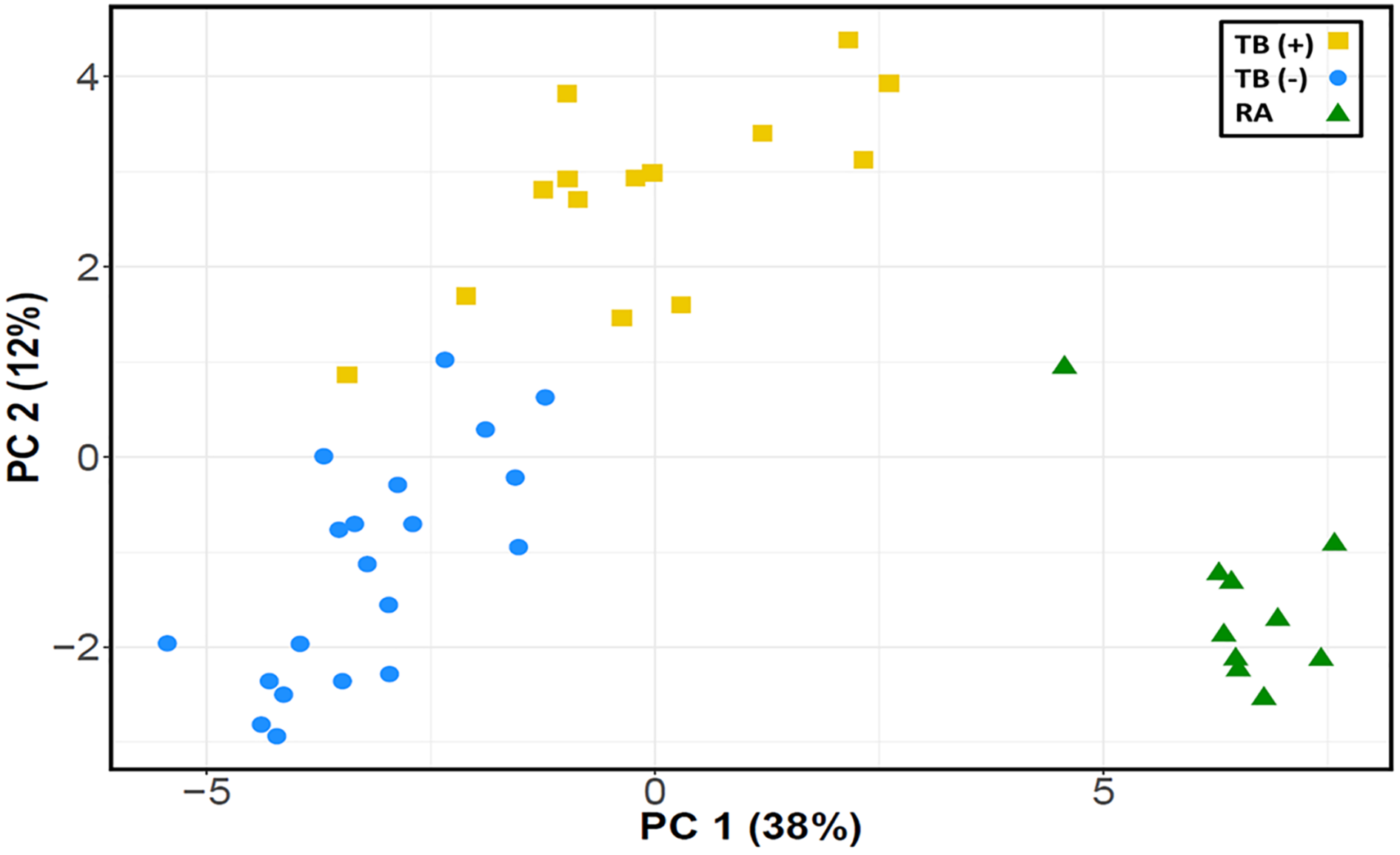


Figure 3

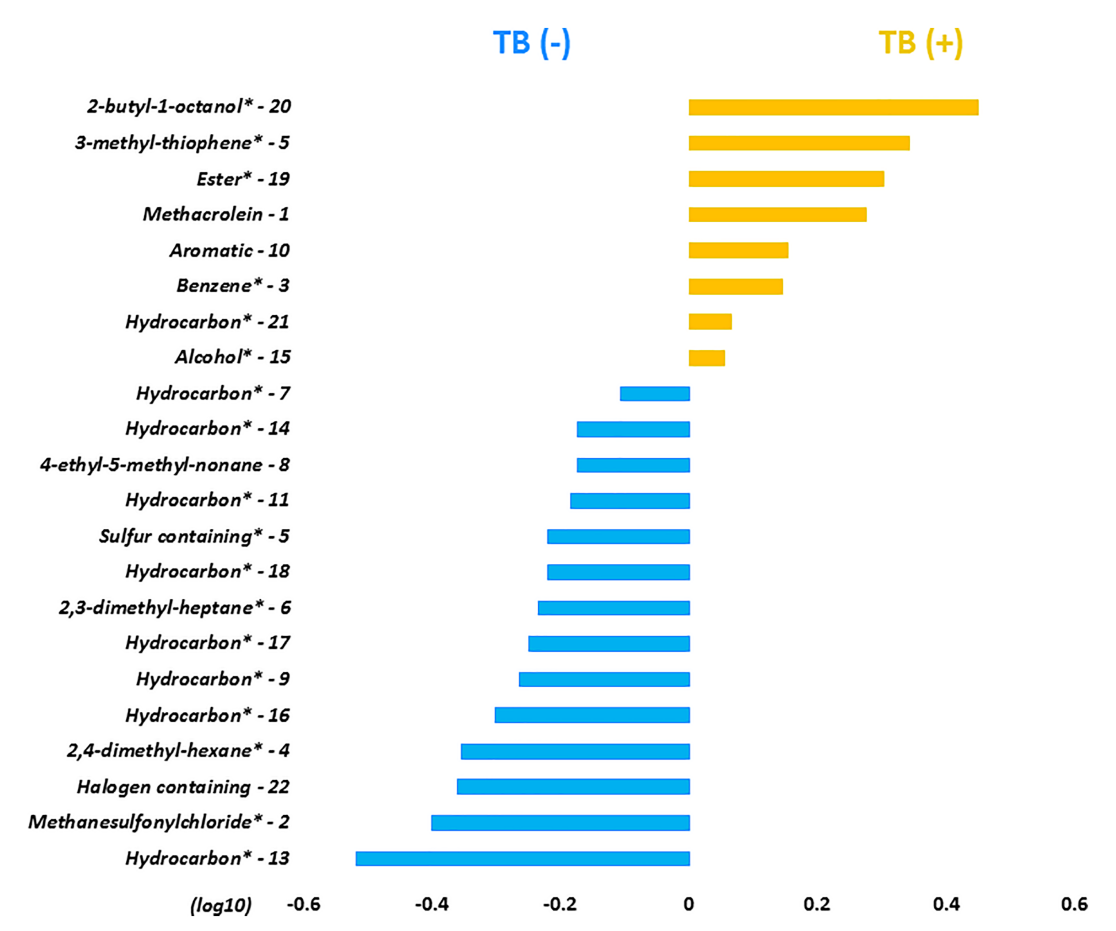

TB (-)
TB (+)
2-butyl-1-octanol* - 20
3-methyl-thiophene* - 5
Ester* - 19
Methacrolein - 1
Aromatic - 10
Benzene* - 3
Hydrocarbon* - 21
Alcohol* - 15
Hydrocarbon* - 7
Hydrocarbon* - 14
4-ethyl-5-methyl-nonane - 8
Hydrocarbon* - 11
Sulfur containing* - 5
Hydrocarbon* - 18
2,3-dimethyl-heptane* - 6
Hydrocarbon* - 17
Hydrocarbon* - 9
Hydrocarbon* - 16
2,4-dimethyl-hexane* - 4
Halogen containing - 22
Methanesulfonylchloride* - 2
Hydrocarbon* - 13
(log10)
-0.6
-0.4
-0.2
0
0.2
0.4
0.6

Figure 4