Journal of Breath Research

Exhaled human breath analysis in active pulmonary tuberculosis diagnostics by comprehensive gas chromatography-mass spectrometry and chemometric techniques

To cite this article before publication: Marco Beccaria *et al* 2018 *J. Breath Res.* in press https://doi.org/10.1088/1752-7163/aae80e

Exhaled human breath analysis in active pulmonary tuberculosis diagnostics by comprehensive gas chromatography – mass spectrometry and chemometric techniques

Marco Beccaria[1,2#], Carly Bobak[3#], Boitumelo Maitshotlo[4], Theodore R. Mellors[1], Giorgia Purcaro[1,5], Flavio A. Franchina[1,6], Christiaan A. Rees[3], Mavra Nasir[3], Andrew Black[7,8], and Jane E. Hill[1,3,*]

[1] Thayer School of Engineering, Dartmouth College, Hanover, NH, 03755, United States

[2] KU Leuven - University of Leuven, Department for Pharmaceutical and Pharmacological Sciences, Leuven, B-3000, Belgium

[3] Geisel School of Medicine, Dartmouth College, Hanover, NH, 03755, United States

[4] Wits Reproductive Health and HIV Institute, Hillbrow, Johannesburg, 2001, South Africa

[5] Gembloux Agro-Bio Tech, University of Liège, Gembloux, 5030, Belgium

[6] Department of Chemistry, University of Liège, Liège (Sart-Tilman), 4000, Belgium

[7] Wits Reproductive Health and HIV Institute University of the Witwatersrand, Johannesburg, 2000, South Africa

[8] Department of Medicine, University of the Witwatersrand, Johannesburg, 2193, South Africa

# These authors contributed equally in this manuscript

[*] Corresponding author. Tel: 1 (603) 646-8656; Fax: 1 (603) 646-8778

E-mail address: Jane.E.Hill@dartmouth.edu

Abstract

Tuberculosis (TB) is the deadliest infectious disease, and yet accurate diagnostics for the disease is unavailable for many sub-populations. In this study, we investigate the possibility of using human breath for the diagnosis of active TB among TB suspect patients, considering also several risk factors for TB as smoker and Human Immunodeficiency Virus (HIV). The analysis of exhaled breath, as an alternative to sputum-dependent tests, has the potential to provide a simple, fast, non-invasive, and ready-available diagnostic service that could positively change TB detection. A total of 50 individuals from a clinic in South Africa were included in this pilot study. Human breath has been investigated in the setting of active TB using thermal desorption-comprehensive two-dimensional gas chromatography–time of flight mass spectrometry methodology and chemometric techniques. From the entire spectrum of volatile metabolites in breath, three machine learning algorithms (Support Vector Machines, Partial Least Squares Discriminant Analysis, and Random Forest) to select discriminatory volatile molecules that could potentially be useful for active TB diagnosis, were employed. Random Forest showed the best overall performance, with sensitivity of 0.82 and 1.00 and specificity of 0.92 and 0.60 in the training and test data respectively. Unsupervised analysis of the compounds implicated by these algorithms suggests that they provide important information to cluster active TB from other patients. These results suggest that developing a non-invasive diagnostic for active TB using patient breath is a potentially rich avenue of research, including among patients with HIV comorbidities.

Keywords: Human exhaled breath; pulmonary Tuberculosis; VOCs; metabolomics; comprehensive two-dimensional gas chromatography; machine learning

1. Introduction

Tuberculosis (TB) is an infectious disease which has been present in humans since ancient times [1]. The disease is caused by the bacterium *Mycobacterium tuberculosis* (Mtb) and primarily infects the lungs (pulmonary TB represents ~85% of TB cases). [1,2]. The World Health Organization estimates that new infections occur in about 1% of the population each year, which in 2016 resulted in more than 10 million cases of active TB. There are several factors that increase the risk of active Mtb infection, such as: malnutrition, tobacco smoking, and several co-pathologies, the most important being co-infection with human immunodeficiency virus (HIV). People living with HIV are anywhere from 26 to 31 times more likely to develop active TB than persons without HIV [3]. Symptoms of active TB disease include at least of one or a combination of the following: cough, fever, night sweats, or weight loss; which are not specifically diagnostic and may be mild for months prior to clinical evaluation.

Diagnosis of pulmonary TB, particularly at primary care level, depends on obtaining an adequate expectorated sputum sample. The gold standard for diagnosis of active TB (bacteriological culture), as well as Nucleic Acid Amplification (NAA) and smear microscopy, are all sputum-dependent. However, up to one third of TB cases cannot reliably produce an adequate biological sputum sample [5]. This can lead to more invasive sampling approaches, including induced sputum or gastric aspirate or a lack of diagnosis altogether, which occurs in many low resource settings. Moreover, risk factors, particularly HIV, can decrease the accuracy of several diagnostic tests, leading to challenges in both the diagnosis and treatment. Therefore, alternative non-invasive samples, such as urine [6**]** and exhaled breath [7] may be useful alternatives of adjuncts in TB diagnosis.

Several research groups, using gas chromatography (GC) linked to mass spectrometry (MS), have investigated the volatile molecules present in breath during Mtb infection in active pulmonary TB, reporting different panels of marker compounds [8-13]. This lack of overlap is likely due to a

multitude of considerations, including: use of different sampling methods and analytical tools as well as patient population heterogeneity, patient co-morbidities (or lack thereof), different control groups, and statistical approaches used. A first step to overpass this lack of standardization was the development of technical standards for breath collection, published recently by Horvath *et al*. [14]. In General, classical clinical parameters, food, drug medications, and smoking habits can also influence breath content. Age and gender may affect breath profiles [15], but their effect are more subtle than smoking behaviors, that can influence the breath profile creating subpopulations [16]. In addition, the profile of Volatile Organic Compounds (VOCs) possibly produced during Mtb infection may be modified by the host at different times during infection [17] and can be variable during the progression/regression of TB disease [12].

In this study, exhaled breath was evaluated from a pilot cohort of 50 patients living in an endemic TB region who were suspected of having TB and includes smokers and subjects with HIV infection. Breath volatile molecules were collected using a multiple-bed sorbent trap and then desorbed, separated, and detected by comprehensive two-dimensional gas chromatography (GC×GC) coupled to a time-of-flight mass spectrometer (TOF MS). Using a variety of machine learning algorithms, we were able to determine volatile metabolic patterns that could be helpful to discriminate between Mtb infected and TB suspect individuals. TB status was confirmed by GeneXpert MTB/RIF® (a NAA test), in combination with bacteriological culture in case of patients with HIV infection.

2. Materials and Methods

2.1 Patient demographics and tuberculosis infection confirmation

A total of 50 individuals, including 32 with active pulmonary TB and 18 controls with TB symptoms, but confirmed Mtb-negative (Johannesburg, South Africa; 2015-2016), were included in the present study. Sputum samples were collected following WHO guidelines for TB [18]. An Institutional

Review Board at the collaborating sites (Wits Reproductive Health and HIV Institute) and Dartmouth approved the research. All subjects gave their signed informed consent to participate and were at least 18 years old. TB status was confirmed by GeneXpert MTB/RIF assay (Cepheid, Sunnyvale, CA, USA). This NAA test is a rapid, automated, cartridge-based test that can detect Mtb along with rifampicin resistance directly from sputum [19]. In individuals with HIV infection, the accuracy of GeneXpert to classify patients with Mtb infection may be unreliable [20], therefore the standard Mycobacteria growth indicator tube (MGIT) bacteriological culture test was employed to confirm Mtb-negative status in HIV-positive subjects (n=4). Patient demographic information is reported in Table 1.

Table 1. Study subject demographic information where n=50.

	Mtb positive (+)	***Mtb negative (-)***	***p-value***
Number (%)	32 (64%)	18 (36%)	0.001
Age, mean (±SD)	35 (±10)	35 (±10)	0.918
Gender (M/F)	18/14	11/7	0.950
Active smoker (Y/N)	7/25	4/14	0.591
HIV (Y/N)	21/11	4/14	0.006
HIV Treatment (Y/N)	8/24	3/15	0.591

2.2 Breath and room air sampling

Prior to breath collection, patients rinsed their mouth with water to avoid some volatile molecule contamination from the oral cavity [21] and then exhaled normally for 2 s into the room [22]. One L Tedlar bags (SKC Inc., Eighty Four, PA, US), pre-conditioned by flushing pure nitrogen gas, were used for the collection of breath over three to five minutes of regular breathing. On the same day of

collection, breath was drawn from the Tedlar bag through a 0.22-μm filter (for the removal of potential pathogens), and onto the thermal desorption tube at a rate of 150 ml/min, for a final breath sampling volume of 1 L. The three-bed thermal desorption (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 [23]. TD tubes containing breath molecules were hermetically sealed and stored at room temperature until further analysis which occurs within a month from collection, as previously reported [13,24-25]. One liter of room air was directly collected into the TD tube on the day of collection.

2.3 Analytical instrumentation

TD tubes were desorbed into a Pegasus 4D (LECO Corporation, St. Joseph, MI) GC×GC-TOF MS instrument with an Agilent 7890 GC equipped with a thermal desorption unit (TDU), cooled injection system (CIS), and a MultiPurpose Sampler (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: 180 s; CIS temperature: 330 °C; injection mode: splitless. Chromatographic analysis was performed using a Rxi-624Sil (60 m × 250 μm × 1.4 μm) as first dimension (1D)- GC column and a Stabilwax (1.5 m × 250 μm× 0.5 μm) as second dimension (2D)-GC column, both purchased from Restek (Bellefonte, PA, US). Modulation time was 2 s total and helium as carrier gas (flowrate: 2 mL/min). TOF MS was employed as detector, 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. Data acquisition and analysis was performed using ChromaTOF software, version 4.50 (LECO Corp.).

2.4 Processing and analysis of chromatographic data

Chromatographic data were processed and aligned using ChromaTOF. For peak identification, a signal-to-noise (S/N) cutoff was set at 150:1 in at least one chromatogram and a minimum of 50:1 S/N ratio in all others. The resulting peaks were identified by a forward search of the NIST 2011 library. For putative peak identification, a forward match score of ≥ 800 (of 1000) was required. For the alignment of peaks across chromatograms, maximum first and second-dimension retention time deviations were set at 6 s and 0.2 s, respectively, and the inter-chromatogram spectral match threshold was set at 600. Compounds eluting prior to 300 s and artifacts (*e.g.,* siloxane, phthalates, etc.) were removed prior to statistical analysis with the support of the script tool available in ChromaTOF®, using the script reported in [26]. An additional data cleaning step was performed to remove common environmental contaminants, artifacts coming from the Tedlar® bag (e.g. phenol and N,N-dimethylacetamide), not included in the script (the complete list of compounds removed is reported in [24]). The most discriminatory features were assigned to a chemical class (Level 3) according to the criteria established by the Metabolomics Standards Initiative (MSI) [27]**,** based on mass spectral similarities to the NIST 2011 mass spectral library, with a match score ≥ 750 (of 1000). Most hydrocarbons were generally assigned as "alkylated hydrocarbons", as it is almost impossible to assign them a specific name based only on the mass spectra similarity, due to the intense fragmentation of this class of compounds into the MS ion source. However, the chemical class of these compounds can be assigned by considering both their location in the two-dimensional chromatogram and their mass spectral fragmentation pattern.

2.5 Statistical analysis

All statistical analyses were performed using R v3.4.3 (R Foundation for Statistical Computing, Vienna, Austria) using "caret" package [28]**.** Prior to statistical analyses, the relative abundance of compounds across chromatograms was normalized using Probabilistic Quotient Normalization [29]

and peak intensities were log-transformed, mean-centered, and then unit-scaled.

Data was randomly subdivided into training (60% of samples) and validation sets (40% of samples). Three machine learning algorithms were used to identify the most discriminatory volatile metabolites and predict the class (Mtb infected versus TB suspect) to which samples in the validation set belonged: Random Forest (RF) [30], Support Vector Machines with a linear kernel (linear SVM) [31], and Partial Least-Squares Discriminant Analysis (PLS-DA) [32]. For each machine learning algorithm, a 5-fold repeated cross validation was employed with 10 repeats [33-34]**.** Mean Decrease in Accuracy (MDA), feature specific Area Under the Receiver Operating Characteristic (AUROC or AUC) curve, and the weighted sums of the absolute regression coefficients were used as the measures of variable importance for RF, linear SVM, and PLS-DA, respectively [28,35]. Features were then selected using the "elbow method" where feature importance was plotted and then a cuttoff was selected in such a way that it captures the "elbow" of the graph. This ensures that any large increases in feature importance were captured and eliminates features which demonstrated only incremental increases in importance [35]. Principal Component Analysis (PCA) [37] was used to visualize the variance between samples in the dataset given our selection of important features. Similarly, Hierarchical clustering analysis (HCA) [38] was used to visualize distance between each sample using Jaccard's distance [39] and a heat map is shown to visualize the relative expression of each feature.

3. Results and Discussion

3.1 Breath evaluation and selected molecules

Contaminants and artifacts (e.g., siloxanes, phthalates) were removed, resulting in a reduction to 1023 features. Moreover, features present in room air sample with a frequency of observation (FOO) ≥50%

were deleted from the matrix, reducing the number of volatile features to 251. At this point, 50% of FOO was applied within the groups to removing sparse features, leading to 128 features which were dominated by hydrocarbons (48%), followed by aromatics (11%), alcohols (8%), halogen-containing compounds (8%), esters (5%), ketones (5%), nitrogen-containing compounds (4%), sulfur-containing compounds (4%), aldehydes (3%), acids (2%), terpenes (1%), and unknowns (1%) (Figure 1b). Prior to any further elaboration the data matrix was normalized using the PQN method, which accounts for dilution of the biological samples. This method uses median values for normalization insuring a stability towards outliers and sampling variability, which can occur in metabolomics [29]. Then, after log-transformation and mean centering, RF, linear SVM, and PLS-DA, were used to identify the most highly discriminatory volatile metabolites from the 128 features list in the discovery set and used to predict the class to which samples in the validation set belonged. A Venn diagram of the panel of 23 features obtained from each machine learning approach is reported in Figure 1c.

<insert Figure 1>

Figure 1. (a) Scheme for feature reduction, (b) chemical class of the 128 features used for data used for statistical elaboration, (c) Venn diagram of the panel of 23 features obtained for the three different machine learning techniques (RF, SVM, and PLS-DA).

Due to the high dimensional nature of -omics data, it is essential that machine algorithms are selected which can handle when the number of features far outweigh the number of samples. Moreover, these algorithms need to also be able to handle highly correlated features (multicollinearity) [40-41]. Practically, feature selection to a manageable size is necessary in order to translate biomarker to a handheld or benchtop system in a clinic or diagnostic laboratory. [42]. RF algorithms generate many classification trees, using randomly selected subsamples of both features and data points. Features

are ultimately selected based on which variables best divides the data according to class at each split [30]. Random Forest has proven to be particularly resilient in -omics classification [40]**.** SVM is a non-parametric method which projects data into some highly dimensional subspace, and then identifies a hyperplane to separate the classes geometrically [43]. The objective of the PLS-DA algorithm is to maximize the covariance between samples and their dependent variable (such as case status) in high dimensional data. To achieve this, it finds a linear sub- space of explanatory variables [44-45]**.** Each of these models has their own sets of parameters which require tuning. To reduce the risk of overfitting, we employed 5-Fold Cross Validation (CV), where our training model was split into 5 approximately even sized pieces, and then we trained the model on 4/5 of these pieces and tested on the remaining piece. We then withhold a separate piece of the data and retrain the model on the remaining 4/5 pieces. We iterate through this process until each piece has been withheld for testing. This allows us to develop an accuracy distribution based on each model's performance on the withheld piece of the data. We repeated our CV scheme 10 times so that multiple different cuts of the training data are considered, thus reducing the variability of the results [46]. We used the entire 5-fold repeated cross validation procedure twice – first to rank our feature importance and apply feature selection, and then again to tune our model parameters used the subset of selected features. The final models after feature selection and tuning were then used on the validation data to evaluate their performance on unseen data.

For each model, we evaluated the accuracy, sensitivity, specificity and AUC in order to assess prediction errors [47]. Table 2 shows each statistic for each of the three final models in both the training and validation datasets. While the performance of all three models on the validation set is strong, both the SVM and PLS-DA models had slightly poorer performance in the validation data. The RF model had similar performance in both the training and validation sets. While the specificity in the validation data is low, this may be partly driven by the low number of 'true negatives' in our

validation set (n=5). The high level of sensitivity in the validation data may indicate that the selected volatile features may be useful in the development of a 'rule-out' TB diagnostic, wherein a negative result from a diagnostic developed from these features would 'rule-out' a TB diagnosis with a high degree of certainty. This would have utility in the clinic as a tool which could be used to screen patients who have a low probability of having TB so they can avoid unnecessary invasive testing using the gold standard diagnostic [3].

Table 2. Accuracy, Sensitivity, Specificity, and AUROC obtained by the machine learning techniques used

	RF		SVM		PLS-DA	
	Training	Validation	Training	Validation	Training	Validation
Accuracy	0.87	0.90	1.00	0.85	0.90	0.80
Sensitivity	0.82	1.00	1.00	0.87	0.94	0.87
Specificity	0.92	0.60	1.00	0.80	0.84	0.60
AUROC	0.93	0.96	1.00	0.89	0.99	0.85

The Receiving Operator Characteristic curves (ROC) for the training and validation sets in each of the three models are shown in Figure 2. The final SVM model had an AUC of 1 in the training data and 0.89 in the withheld validation data. The final PLS-DA model had an AUC of 0.99 in the training data, and 0.85 in the validation data. The final RF model had an AUC of 0.93 in the training data, and 0.96 in the validation data. Given the RF models superior performance in the withheld validation data, we selected it as the best model for classification of active TB patients in this particular data set.

<insert Figure 2>

Figure 2. Receiver (or Relative) Operating Characteristic (ROC) Curve by using SVM, PLS-DA, and

RF algorithms. For each machine learning technique, the set of molecules generated in Training set (n=30) were tested in the Validation set (n=20).

To assess whether a bias due to class imbalance was present due to the limited number of HIV-/Mtb+ samples, the accuracy of the model within this particular subgroup of data was evaluated. Overall, the RF model classified 75% of this group correctly, and hence we do not think class imbalance greatly affected our results.

There was significant overlap of selected features across our three models. In total, 23 features were selected in total from all three models, with 12 features in common to all models. The high conservation of volatile features across these three disparate models increases our confidence that these features are potentially discriminatory molecules for active TB diagnosis on this study population. In Table 3, the rank of each feature for each model is given for the three machine learning techniques, the match of the feature with the NIST library, and retention time of each feature in the first and second dimensions are reported. More than 60% of volatile metabolites detected can be attributed to chemical classes related to the lipid oxidation pathways, namely ketones, aldeheydes, alcohols, and in paricular hydrocarbons (around 50%). These sorts of molecules have been reported to originate largely from free radical oxidative fragmentation of lipids due to oxidative stress [48].

To visualize the ability of these features to discriminate active TB among TB suspects, we used an HCA and PCA developed using all 23 features selected by any of the three models which is shown in Figure 3.

<insert Figure 3>

Figure 3. A Heatmap showing the unsupervised clustering of all 23 features discovered across the three machine learning techniques (RF, SVM, and PLS-DA).

Table 3. Machine learning model feature ranking and analytical context

<Insert Table 3>

The HCA and subsequent heatmap shown in Figure 3 shows the HCA analysis where the distance between samples was calculated using Jaccard's Index, a distance metric which has previously shown to be resilient to noise [39]. All features selected by any of our models are shown on the vertical axis while the unique patient number and their TB and HIV status are shown on the horizontal axis. Notably, as seen by the blue and yellow annotation bar, all of the TB+ cases cluster together, with only 2 out of 14 of the TB-/HIV- cases clustering away from the TB- group. Of note, both of these cases are HIV-, which indicates that these cases are not clustering away from the other TB- cases due to confounding by HIV status. Hence, we believe that the volatile biomarkers selected by our algorithms are not sensitive to HIV status.

<insert Figure 4>

Figure 4. (a) PCA of the 23 discriminatory features obtained after 3 different machine learning techniques (RF, SVM, and PLS-DA). (b) Boxplot showing the first PC component score for each of the TB/HIV subgroups of interest, as well as a global Kruskall-Wallis p-value. Two-way comparisons between TB+/TB- subgroups are also shown, where the number of stars indicate the significance of a Wilcoxon rank-sum test.

A PCA developed using all 23 selected features is shown in Figure 4a, where the color maps to TB/HIV case status (blue is TB-, yellow is TB+, while the darker shades are HIV- and the bright shades HIV+). While we do not observe distinct clusters by case status, a general assortment of TB-

cases to TB+ cases along the PC1 axis is observed. To further examine this effect, we examine the distribution of the PC1 scores across the TB/HIV sub-groups of interest using a boxplot in Figure 4b. We can clearly see differences between the TB+/TB- patients by the PC1 score. A global Kruskal-Wallis test rejected the hypothesis that these samples originated from the same distribution with a highly significant p-value of $9.1e^{-7}$. We also conducted to two-way comparisons between the various TB+/TB- subgroups using Wilcoxon's rank-sum test. All comparisons were significant at a Benjamini-Hochberg corrected significant level of $\alpha = 0.05$. With additional samples, we expect this effect to become clearer.

Similar behavior was observed using the discriminatory features obtained after cross validation considering the single machine learning technique applied (14 for RF, 21 for SVM, and 17 for PLS-DA), but also considering the 12 common features within each model (Figure 1c). HCA and PCA plots for each machine learning model utilized in our analyses are available in the supplementary data (Figure S1-S3), while Figure S4 shows HCA and PCA plots of the 12 common features for each model

3.2. Study strengths and limitations

In the present, pilot study, we evaluated the potential ability of volatile molecules in the breath for discriminating between Mtb-infected and TB-suspect individuals using three different machine learning algorithms. Twenty-three discriminatory features were selected using the different algorithms (PLS-DA, SVM, and RF). Although a good match with the library was obtained (20 out of 23 features had a match > 800/1000 and the other 3 > 750/1000), we preferred to not report a putative identification of these possible biomarkers, since a large cohort study is necessary to validate the biomarkers. Future studies should include a greater proportion of patients who TB suspects that end up being negative for Mtb infection, but who are also co-infected with HIV, as well as a higher

number of co-infected subjects. In addition, other co-morbidities in the patient population e.g., diabetes, would also assist in generating universal biomarkers. Despite the limitations, we plan to evaluate the panel of 23 breath molecules in future studies and hopefully confirmed and validated as biomarkers by using an external dataset. It is important to highlight that the percentage of chemical classes of the 23 breath molecules reported as discriminatory in this pilot study (Table 3) is in according with previous GC based techniques studies on human exhaled breath in the setting of TB disease [8-13].

4. Conclusion

This pilot study (n = 50) is part of a larger, ongoing TB breath biomarker initiative. Here, we demonstrated that volatile metabolites present in human exhaled breath can also be used to discriminate between individual with a positive Mtb infection and people with one or more TB symptoms, but with a confirmed negative Mtb infection. In the validation set, accuracy value was about 0.8-0.9 for all the three machine learning techniques applied, with an AUROC between 0.85 (PLS-DA), and 0.96 (RF). Although all three models showed great prediction power to discriminate those infected with Mtb and TB suspect individuals, the RF model was the most consistent, showing similar performance in both the training and validation sets. This study, along with others, reiterate that exhaled human breath in diseased individuals contains useful data which should be developed as a non-invasive clinical tool to be deployed in efforts to curb the spread of Mtb infection.

Acknowledgements

Dartmouth College holds an Institutional Program Unifying Population and Laboratory Based Sciences award from the Burroughs Wellcome Fund, and C. Bobak and Mavra Nasir were supported by this grant (Grant#1014106). Christiaan A. Rees was supported by the National Institutes of Health

training grant (Grant # T32LM012204).

References

[1] Lawn S D, Zumla A I 2011 Tuberculosis *Lancet* **378** 57–72

[2] Southwick F 2007. "Chapter 4: Pulmonary Infections". *Infectious Diseases: A Clinical Short Course*, 2nd ed. McGraw-Hill Medical Publishing Division. pp. 104, 313–4

[3] World Health Organization (WHO), Global Tuberculosis Report, (2017)

[4] http://www.who.int/tb/areas-of-work/tb-hiv/en/

[5] Parsons L M, Somoskövi Á, Gutierrez C, Lee E, Paramasivan C N, Abimiku A, Spector S, Roscigno G, Nkengasong J 2011. Laboratory Diagnosis of Tuberculosis in Resource-Poor Countries: Challenges and Opportunities. *Clin Microbiol Rev* **24** 314–350

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

[7] Boots A W, van Berkel J J B N, Dallinga J W, Smolinska A, Wouters E F, van Schooten F J 2012 The versatile use of exhaled volatile organic compounds in human health and disease, *J. Breath Res.* **6** 027108

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

[9] Phillips M, Basa-Dalay V, Bothamley G, Cataneo R N, Lam P K, Natividad M P R, Schmitt P, Wai J 2010 Breath biomarkers of active pulmonary tuberculosis, *Tuberculosis* **90** 145–151.

[10] Phillips M et al 2012 Point-of-care breath test for biomarkers of active pulmonary tuberculosis *Tuberculosis* **92** 314–320.

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

777–782.

[12] Dang N A, Janssen H G, Kolk A H 2013 Rapid diagnosis of TB using GC–MS and chemometrics, *Bioanalysis* **5** 3079–3097.

[13] Beccaria M et al 2018 Preliminary investigation of human exhaled breath for tuberculosis diagnosis by multidimensional gas chromatography – Time of flight mass spectrometry and machine learning *J Chromatogr. B* **1074-1075** 46-50

[14] Horváth I *et al* 2017 A European Respiratory Society technical standard: exhaled biomarkers in lung disease Eur Respir J **49** 1600965

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

[16] Blanchet L, Smolinska A, Baranska A, Tigchelaar E, Swertz M, Zhernakova A, Dallinga J W, Wijmenga C, van Schooten F J 2017 Factors that influence the volatile organic compound content in human breath *J Breath Res* **11** 016013

[17] Bean H, Jiménez-Díaz J, Zhu J, Hill J E 2015 Breathprints of model murine bacterial lung infections are linked with immune response *Eur. Respir. J.* **45** 181–190

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

[19] World Health Organization. Automated real-time nucleic acid amplification technology for rapid and simultaneous detection of tuberculosis and rifampicin resistance: Xpert MTB/RIF assay for the diagnosis of pulmonary and extrapulmonary TB in adults and children: policy update [Internet] Geneva: World Health Organization; 2013. [cited 2015 Mar 1, Available from:http://www.who.int/iris/handle/10665/112472#sthash.WDSfafG9.dpuf.

[20] Lawn S D *et al* 2013 Advances in tuberculosis diagnostics: the Xpert MTB/RIF assay and future

prospects for a point-of-care test *Lancet Infect. Dis.* **13** 349-61

[21] Caddy G R, Sobell M B, Sobell L C 1978 Alcohol breath tests: Criterion times for avoiding contamination by "mouth alcohol" *Behav Res Methods Instrum* **10** 814-818

[22] Mochalski P, Wzorek B, Sliwka I., Amann A 2009 Suitability of different polymer bags for storage of volatile sulphur compounds relevant to breath analysis *J. Chromatogr. B* **877** 189–196

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

[24] Mellors T R, Blanchet L, Flynn J L, Tomko J, O'Malley M, Scanga C A, Lin P L and Hill J E 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* 1**22** 695-701

[25] Mellors T R *et al* 2018 Identification of Mycobacterium tuberculosis using volatile biomarkers in culture and exhaled breath *J. Breath Res.* in Press. DOI: 10.1088/1752-7163/aacd18

[26] Purcaro G, Stefanuto P, Franchina F. A, Beccaria M, Wieland-Alter W F, Wright P F, Hill J E 2018 Fingerprint of cell cultute infected by virus: sample preparation optimization and data processing evaluation *Anal Chim Acta* **1027** 158-167

[27] Sumner L W, Samuel T, Noble R, Gmbh S D, Barrett D, Beale M H, Hardy N 2007 Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI) *Metabolomics* **3** 211–221

[28] Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, ZacharyMayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan and Tyler Hunt. (2018). caret: Classification and Regression Training. R package version 6.0-79 https://CRAN.R-project.org/package=caret

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

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

[31] Cortes C and Vapnik V 1995 Support-Vector Networks *Mach Learn* **20** 273–97

[32] Barker M and Rayens W 2003 Partial least squares for discrimination *J Chemom* **17** 166–73

[33] Mosteller F , Tukey JW 1968 Data analysis, including statistics. In Handbook of Social Psychology. Addison-Wesley, Reading, MA, 1968.,

[34] Kohavi R 1995 A study of cross-validation and bootstrap for accuracy estimation and model selection *In Ijcai* **14** 1137-1145

[35] Krooshof P W T, Ustun B, Postma G J and Buydens L M C 2010 Visualization and recovery of the (Bio)chemical interesting variables in data analysis with support vector machine classification *Anal Chem* **82** 7000–7.

[36] Brieuc M S, Waters C D, Drinan D P, Naish K A 2018 A practical introduction to random forest for genetic association studies in ecology and evolution. *Mol Ecol Resour* **18** 755-766

[37] Hotelling H 1933 Analysis of a complex of statistical variables into principal components *J Educat Psychol* **24**(6), 417.

[38] Tibshirani R, Friedman J 2001 The elements of statistical learning: data mining, inference, and prediction. Heidelberg: Springer.

[39] Toldo R, Fusiello A 2008 Robust multiple struc- tures estimation with j-linkage. In Lecture Notes in Computer Science, pages 537–547. Springer Berlin Heidelberg.

[40] Lebedev A *et al* 2014 Random Forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness *NeuroImage: Clinical* **6**:115–125

[41] Statnikov A, Wang L, Aliferis C F 2008 A comprehensive comparison of random forests and

support vector machines for microarray-based cancer classification *BMC bioinformatics* **9**(1), 319

[42] Pal N R 2007 A fuzzy rule based approach to identify biomarkers for diagnostic classification of cancers. In Fuzzy Systems Conference. FUZZ-IEEE 2007. IEEE International (pp. 1-6). IEEE.

[43] Suykens J A, Vandewalle. J 1999 Least squares support vector machine classifiers *Neural Processing Lett* **9**(3) 293–300

[44] Gromski P S, Muhamadali H, Ellis D I, Xu Y, Correa E, Turner M. L, Goodacre R 2015 A tutorial review: Metabolomics and partial least squares-discriminant analysis – a marriage of convenience or a shotgun wedding *Anal Chim Acta* **879** 10–23.

[45] Pérez-Enciso M, Tenenhaus M 2003 Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (PLS-DA) approach *Human genetics* **112** 581–592

[46] Mosteller F, Tukey J W 1995 Data analysis, including statistics. In Handbook of Social Psychology. Addison-Wesley, Reading, MA, 1968

[47] Fielding A, Bell J 1997 A review of methods for the assessment of prediction errors in conservation presence/absence models *Environ Conserv* **24**(1) 38-49

[48] Schulz S and Dickschat J S 2007 Bacterial volatiles: the smell of small organisms Nat. Prod. Rep. 24 814–42; Haick H, Broza Y Y, Mochalski P, Ruzsanyi V and Amann A 2014 Assessment, origin, and implementation of breath volatile cancer markers *Chem Soc Rev* **43** 1423–49

Figure Captions

Figure 1. (a) Scheme for feature reduction, (b) chemical class of the 128 features used for data used for statistical elaboration, (c) Venn diagram of the panel of 23 features obtained for the three different machine learning techniques (RF, SVM, and PLS-DA).

Figure 2. Receiver (or Relative) Operating Characteristic (ROC) Curve by using SVM, PLS-DA, and RF algorithms. For each machine learning technique, the set of molecules generated in Training set (n=30) were tested in the Validation set (n=20).

Figure 3. A Heatmap showing the unsupervised clustering of all 23 features discovered across the three machine learning techniques (RF, SVM, and PLS-DA).

Figure 4. (a) PCA of the 23 discriminatory features obtained after 3 different machine learning techniques (RF, SVM, and PLS-DA). (b) Boxplot showing the first PC component score for each of the TB/HIV subgroups of interest, as well as a global Kruskall-Wallis p-value. Two-way comparisons between TB+/TB- subgroups are also shown, where the number of stars indicate the significance of a Wilcoxon rank-sum test.

Table Captions

Table 1. Study subject demographic information

Table 2. Accuracy, Sensitivity, Specificity, and AUROC obtained by the machine learning techniques used

Table 3. Machine learning model feature ranking and analytical context

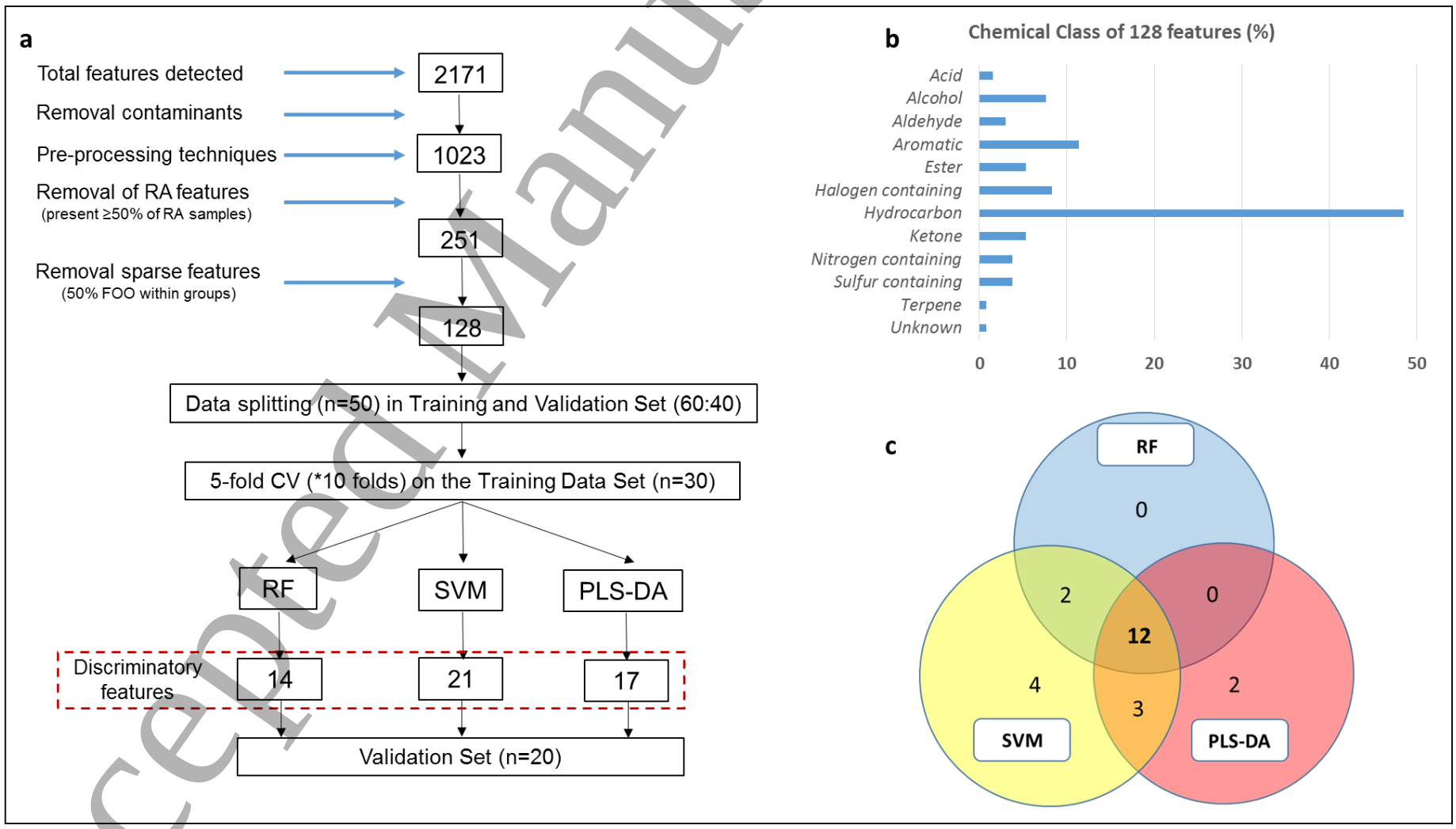

Figure 1

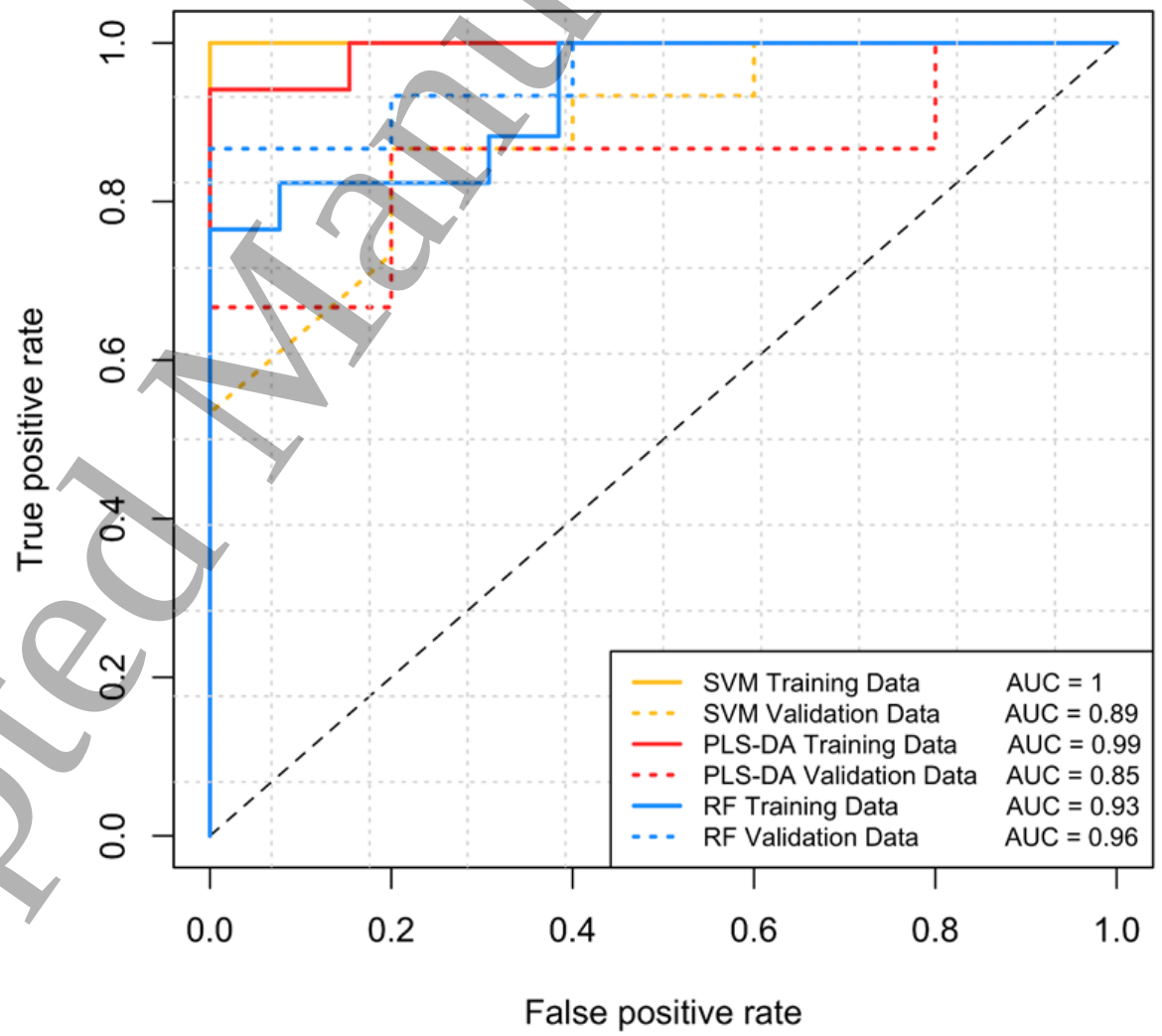

Figure 2

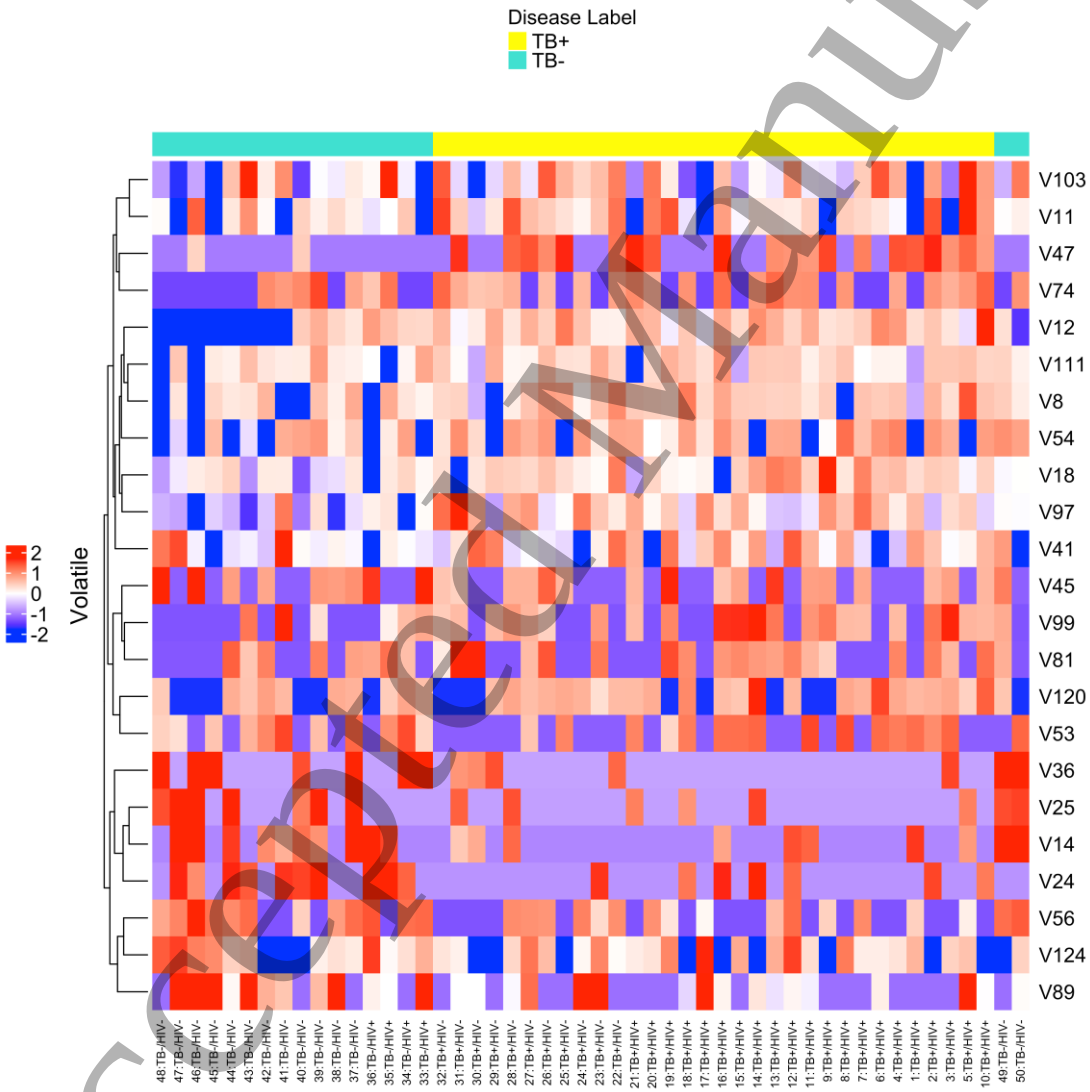

Figure 3

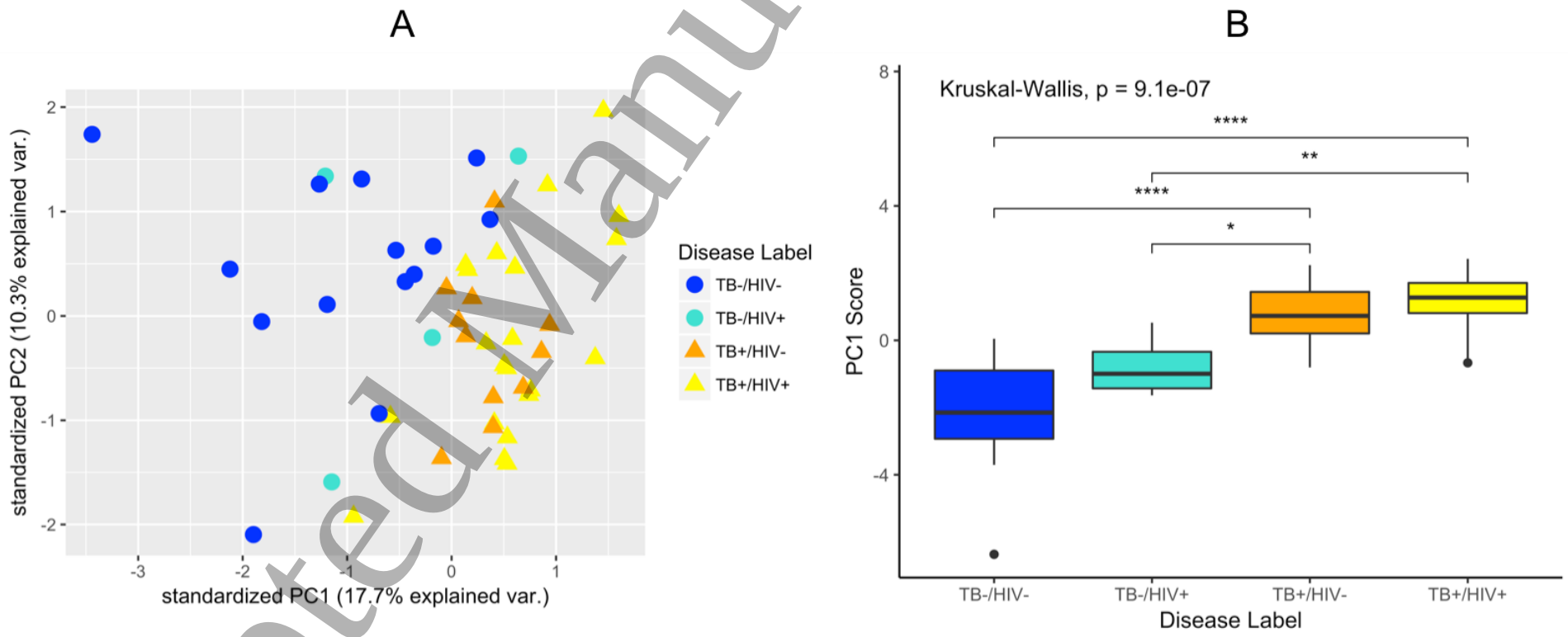

Figure 4

Table 3. Machine learning model feature ranking and analytical context

# Feature	RF	SVM	PLS-DA	Average score	Chemical class	Forward similarity	Reverse similarity	$^{1}t_{R\ (sec)}$	$^{2}t_{R\ (sec)}$
47	1	2	1	1.33	Alkylated hydrocarbon	823	847	1876	0.59
97	2	1	2	1.67	Halogen containing	888	888	1738	0.65
103	3	3	4	3.33	Aldehyde	903	903	1466	0.67
12	4	14	3	7.00	Halogen containing	874	874	2561	0.60
36	5	11	7	7.67	Hydrocarbon	842	856	1136	0.60
8	6	6	11	7.67	Alkylated hydrocarbon	877	877	2115	0.60
99	14	4	6	8.00	Alkylated hydrocarbon	921	921	1136	0.57
11	10	5	10	8.33	Acid	753	778	1356	0.64
25	12	9	5	8.67	Alkylated aromatic	850	850	708	0.67
56	8	8	15	10.33	Alkylated hydrocarbon	860	860	2583	0.61
14	7	13	12	10.67	Alkylated hydrocarbon	865	865	804	0.62
89	13	12	14	13.00	Aldehyde	809	867	2418	0.66
54	19	15	9	14.33	Alkylated hydrocarbon	811	827	2386	0.60
53	26	10	8	14.67	Alkylated hydrocarbon	876	876	2742	0.61
124	22	16	18	18.67	Alkylated hydrocarbon	900	900	1507	0.58
74	18	26	13	19.00	Alkylated ester	821	821	1978	1.80
18	9	7	42	19.33	Alkylated alcohol	838	838	2357	0.60
41	20	17	22	19.67	Alkylated hydrocarbon	774	879	1443	0.58
24	27	19	16	20.67	Alkylated hydrocarbon	775	800	1848	0.59
45	16	20	32	22.67	Alkylated hydrocarbon	862	911	1336	0.59
111	11	18	43	24.00	Ester	926	926	1910	0.69
120	29	44	17	30.00	Alkylated alcohol	828	828	2722	0.60
81	54	21	25	33.33	Cyclo-alcohol	881	886	1460	0.80