Exhaled Breath Analysis: A Tool For Exposomics – Determining Smoking Status Using VOC Biomarkers

Stefano Patassini[1], Mike O'Neill[1], David Layton[1], Elisabeth Krizek[1], Giovanna De Palo[1], Megan Williams[1], Yichen [Kelly] Chen[1], Theodore W. Wilson[1], William Murch[1] and Max Allsworth[1]

[1]Owlstone Medical Ltd., Unit 183, Cambridge Science Park, Milton Road, Cambridge, CB4 0GJ

1.0 Introduction

Exhaled breath contains thousands of volatile organic compounds (VOCs) originating as products of the body's metabolism (endogenous VOCs) or from external sources such as diet, occupational and environmental exposure (exogenous VOCs). Breath VOCs therefore provide valuable, non-invasive information about the volatile fraction of both the internal and external exposomes – e.g. VOCs inhaled during a shift at a factory, or volatile metabolites from processes related to an exposure such as inflammation of lung tissues as the result of particle inhalation or oxidative stress.

To demonstrate how breath can be used to discover breath-based biomarkers of exposure, this study used the Breath Biopsy Platform to collect and analyse breath samples from smokers and non-smokers to identify smoking related VOC biomarkers.

2.0 Methods

Exhaled breath samples were collected using the ReCIVA breath sampler (Figure 1). For each tube, 1.473 L breath was collected over a ~10-minute period using a ReCIVA connected to a CASPER Portable Air Supply. Tubes were shipped to Owlstone Medical's Breath Biopsy Laboratory for analysis. Samples were pre-purged to remove excess water and desorbed using a TD100-xr thermal desorption autosampler (Markes International) and transferred onto a GC column (Agilent Technologies).

Chromatographic separation was achieved via a programmed method on a Trace 1310 GC oven (Thermo Fisher Scientific) and mass spectral data acquired using an electron ionisation time-of-flight BenchTOF HD mass spectrometer (Markes International). Raw data files were converted using TOF-DS (Markes International) and MassHunter (Agilent Technologies) was used for peak area extraction. Measured spectra for relevant features were compared against the NIST unit mass spectral library in order to assign a tentative ID.

Figure 1. The ReCIVA Breath Sampler

3.0 Results

Analysis of breath samples from 73 subjects (17 smokers and 56 non-smokers) yielded 475 molecular features (MFs) with distinct mass spectrums and retention times were identified across breath samples, 26 MFs show a statistically significant fold change following Bonferroni correction between the two classes (non-smokers vs. smokers). Analysis revealed 26 MFs that were significantly different between the groups. The statistically significant features were analysed by quantifying how well the two classes separated from each other in the dimension of the feature across all samples – yielding receiver operating characteristics-area under curves (ROC-AUCs) ranging between 0.72 and 0.96. The top 25 features ranked by p-value are listed in Table 1 and box plots of distribution of peak area measured in non-smokers vs. smokers for each feature are shown in Figure 2. Significant p-values were only found in features with negative fold changes (increased in smokers compared to non-smokers).

Feature	log2 fold change	*p*-value	Smokers Mean Peak Area	Non-smokers Mean Peak Area	STDEV	Tentative ID	Classification Performance (AUC)
MF58	-2.502	2.57E-20	116,346,866	20,545,985	48,679,811	Benzene	0.966
MF57	-1.117	4.79E-15	216,630,371	99,856,608	65,204,476	Fumaronitrile	0.96
MF38	-2.587	3.32E-14	49,496,511	8,238,056	23,513,523	Cyclopentane, methyl-	0.885
MF44	-4.012	1.24E-11	16,986,930	1,053,196	9,804,014	2-Butynoic acid	0.939
MF53	-4.130	1.49E-11	21,379,987	1,221,323	12,437,991	2-Pentyn-4-one	0.913
MF115	-2.045	3.77E-11	109,944,393	26,643,174	52,154,849	Benzene, 1,3-dimethyl-	0.855
MF19	-2.135	3.81E-11	49,831,002	11,342,495	24,100,986	Oxetane, 3-(1-methylethyl)-	0.933
MF113	-2.325	5.92E-11	97,550,305	21,383,311	49,923,661	Ethylbenzene	0.845
MF124	-2.437	2.79E-10	61,342,065	11,327,584	32,408,012	Styrene	0.825
MF93	-6.098	6.65E-10	38,165,966	567,324	24,918,772	Furan, 2-ethyl-5-methyl-	0.918
MF63	-4.692	1.19E-08	36,681,912	1,419,053	24,698,364	1H-Imidazole-4-ethanamine, .beta.-methyl-	0.878
MF86	-1.072	4.88E-08	217,848,071	105,413,331	82,062,878	Toluene	0.862
MF75	-4.852	6.06E-08	8,501,830	342,744	6,296,365	Phenol	0.866
MF9	-1.328	1.26E-07	16,327,469	6,501,395	7,313,154	2-Butyne	0.784
MF125	-2.468	2.47E-07	6,695,671	1,231,925	4,182,140	1-Methyltricyclo[2.2.1.0(2,6)]heptane	0.822
MF142	-1.865	6.06E-07	13,907,661	3,816,871	7,874,703	2-Propanamine, N,N'-methanetetraylbis-	0.817
MF162	-1.281	1.08E-06	25,870,737	10,648,178	12,108,313	Ethane, 1,1,1-trifluoro-	0.824
MF21	-1.507	2.91E-06	30,603,774	10,766,160	16,348,183	Oxetane, 3-(1-methylethyl)-	0.863
MF31	-1.246	2.95E-06	106,454,692	44,895,665	50,756,357	Methyl isocyanide	0.801
MF158	-1.763	3.10E-06	49,333,978	14,532,551	28,746,786	Benzene, 1,2,3-trimethyl-	0.785
MF126	-2.901	5.01E-06	17,271,925	2,353,880	12,658,290	2-(Butyliden-2-one)tetrahydrofuran	0.796
MF146	-3.275	1.01E-05	20,261,818	2,093,103	15,733,426	Propylphosphonic acid, fluoroanhydride, propyl ester	0.827
MF153	-1.818	1.11E-05	45,675,238	15,745,134	29,350,726	Benzene, 1-ethyl-4-methyl-	0.723
MF123	-1.606	2.25E-05	48,690,334	19,264,481	30,065,201	Benzene, 1,3-dimethyl-	0.736
MF43	-3.641	4.47E-05	9,050,245	725,273	7,723,706	1-(1'-pyrrolidinyl)-2-butanone	0.869

Table 1. Top 25 molecular features identified in breath samples ranked by p-value

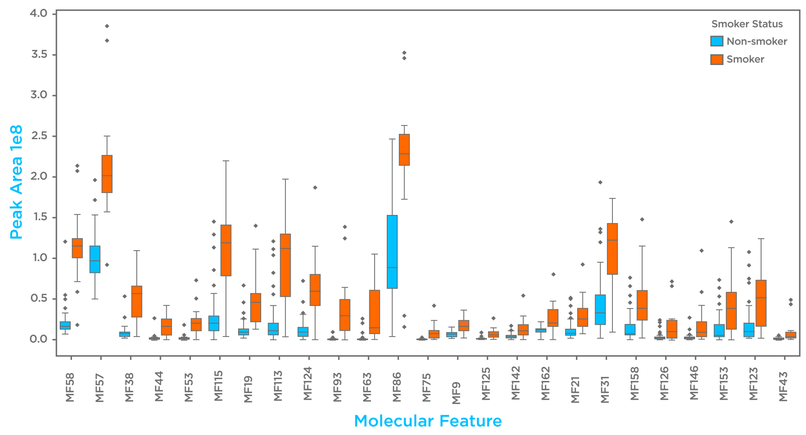
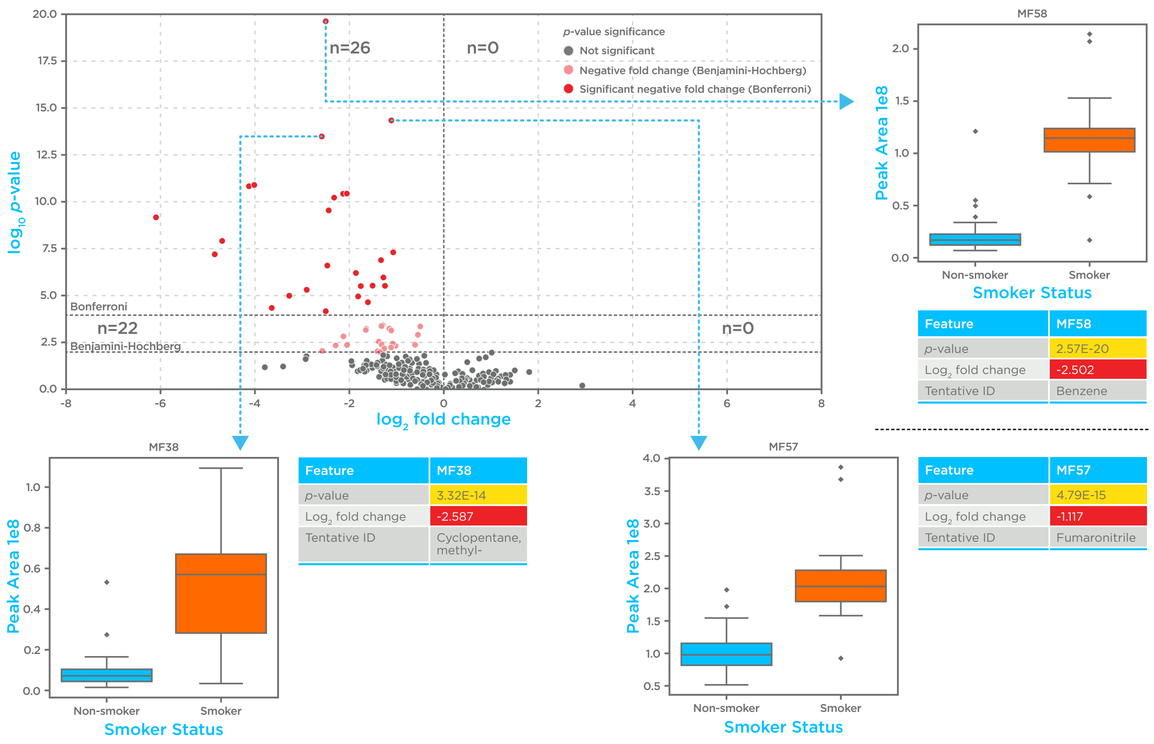


Figure 2. Box plots of peak area for top 25 molecular features ranked by p-value (left to right).

On the volcano plot (Figure 3), MFs with significant p-values above the Bonferroni cut-off are shown in full red, those with p-values above the Benjamini-Hochberg cut-off are shown in partial red. The box plots for highlighted MFs (lower panels) show the distribution of peak area measured for each feature in non-smokers vs. smokers. Each table shows p-value, $\log_2$ fold change between classes (non-smokers vs. smokers) and tentative ID for the MF.



Feature	MF58
p-value	2.57E-20
Log_2 fold change	-2.502
Tentative ID	Benzene

Feature	MF38
p-value	3.32E-14
Log_2 fold change	-2.587
Tentative ID	Cyclopentane, methyl-

Feature	MF57
p-value	4.79E-15
Log_2 fold change	-1.117
Tentative ID	Fumaronitrile

Figure 3. Volcano plot of all identified MFs, box plots of peak area distribution included for highlighted MFs showing fold change between classes (non-smokers vs smokers)

3.1 Building Classifier Using Random Forest

Random Forest is a supervised machine learning algorithm used for classifier building. In this case a combination of MFs discriminated between smokers and non-smokers with ROC-AUC = 0.97 – using 10-fold cross-validation (Figure 4).

A confusion matrix is shown in Figure 5 describing the performance of the Random Forest Classification Model to classify the samples (top), prediction probabilities of individual samples (bottom left), and box plots of each class (bottom right). Dashed line represents the Non-smoker threshold (0.5).

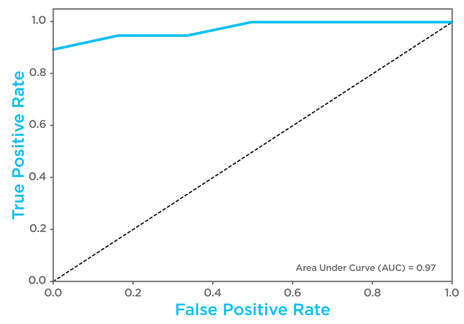


Figure 4. ROC using Random forest classification

Subject classification at Non-smoker threshold 0.5

		Breath Test (Predicted Label)		
		Smoker	Non-smoker	Total
True label	Smoker	15 True Negatives	2 False Positives	17
	Non-smoker	1 False Negatives	55 True Positives	56
Total		16	57	73

Performance of the Classifier

Sensitivity (55/56)	98%
Specificity (15/17)	88%
Positive Predictive Value (55/57)	96%
Negative Predictive Value (15/16)	94%

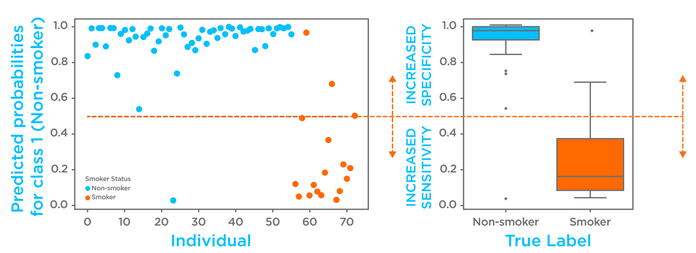


Figure 5. Confusion matrix (top), prediction probabilities of individual samples (bottom left) for random forest classification

3.2 Building Classifier Using LDA

Linear Discriminant Analysis (LDA) is a machine learning technique that creates a series of discriminant functions comprised of linear combinations of features. These functions maximise the distances between the two classes. Utilizing these functions LDA will then classify the samples into one group or another. In this case a combination of MFs discriminated between smokers and non-smokers with ROC-AUC = 0.96, also using 10-fold cross-validation (Figure 6).

Figure 7 shows a confusion matrix describing the performance of the LDA Model to classify the samples (top). The figures describe the prediction probabilities of individual samples (bottom left), and box plots of each class (bottom right). Dashed line represents the Non-smoker threshold (0.5).

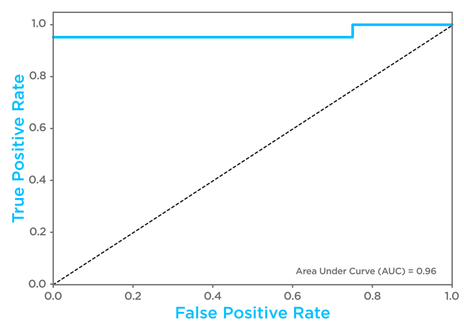


Figure 6. ROC using LDA

Subject classification at Non-smoker threshold 0.5

		Breath Test (Predicted Label)		
		Smoker	Non-smoker	Total
True label	Smoker	14 True Negatives	3 False Positives	17
	Non-smoker	2 False Negatives	54 True Positives	56
Total		16	57	73

Performance of the Classifier

Sensitivity (55/56)	98%
Specificity (15/17)	88%
Positive Predictive Value (55/57)	96%
Negative Predictive Value (15/16)	94%

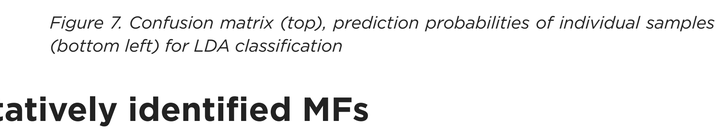


Figure 7. Confusion matrix (top), prediction probabilities of individual samples (bottom left) for LDA classification

3.3 Quantification of Tentatively identified MFs

To verify the validity of our findings, some tentatively identified MFs underwent further quantification using pure synthetic standards as VOC surrogates. Calibration curves for tentative MF compounds were constructed (example benzene curve shown in Figure 8a). Quantification of 6 compounds in 135 samples showed those compounds to be present in breath at parts per billion level (Figure 8b). It was found that the concentration of some compounds was significantly higher in the breath of current smokers compared to individuals who have never smoked/given up smoking. This is demonstrated in example box plots (Figure 8b) for BTEX (benzene, toluene, ethylbenzene and p-xylene) and 2-methylfuran.

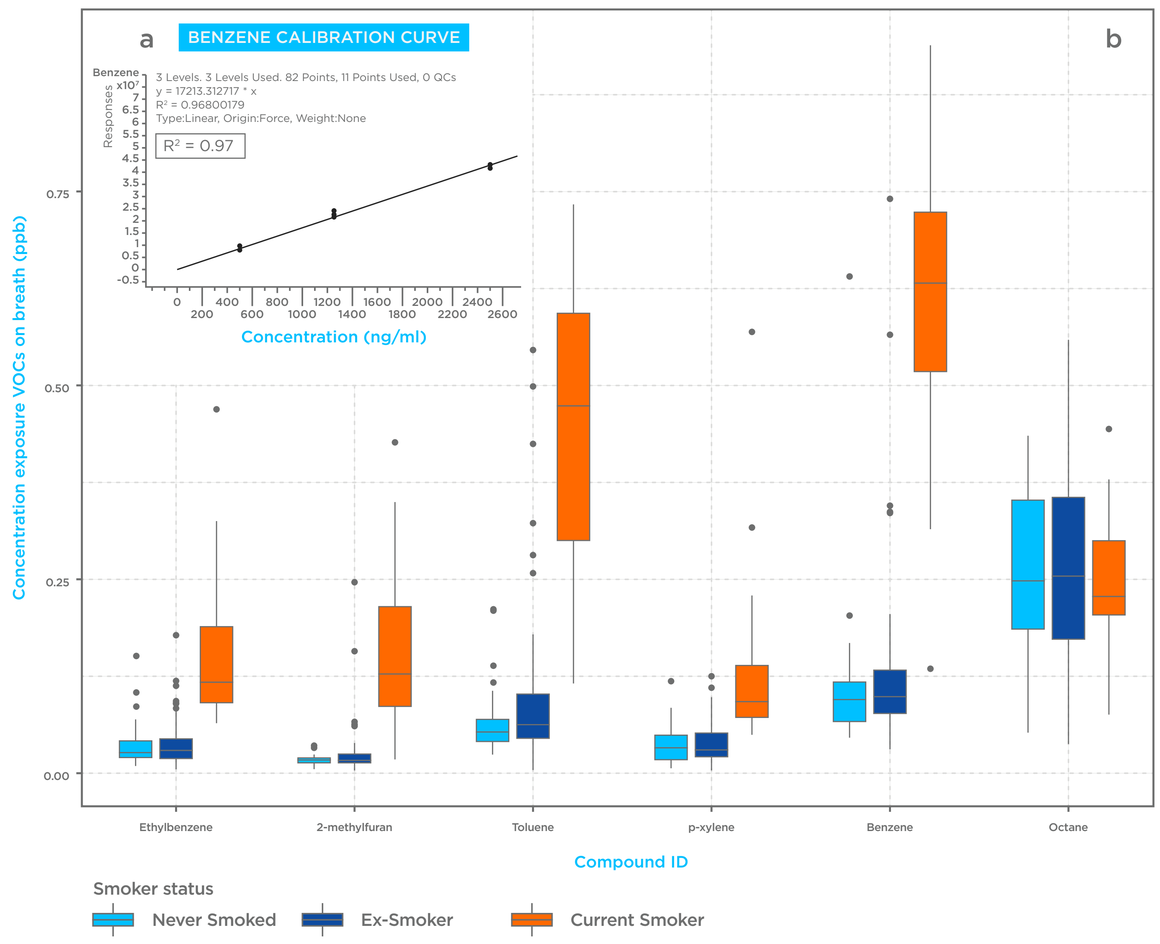


Figure 8. Quantification of selected VOCs compounds in breath samples.
a) Example calibration curve and b) boxplots showing differences in breath concentration of multiple compounds between never-smoked, ex-smokers and current smokers.

4.0 Conclusions

Analysis of 73 breath samples revealed 26 molecular features (MFs) significantly different between smokers and non-smokers. Quantifying separation between classes in the dimension of the feature yielded ROC-AUCs ranging between 0.72 and 0.96. Combinations of MFs analysed using LDA and random forest discriminated between the groups with ROC-AUCs of 0.96 and 0.97 respectively – using 10-fold cross-validation. Tentative molecular identification of MFs indicated many are common combustion related compounds, e.g. BTEX.

Using breath samples collected from 136 individuals, a subset of the MFs were further quantified at the PPB level, which confirmed the ability of BTEX compounds and 2-methylfuran to discriminate between smokers and non-smokers. This study supports breath analysis as a novel technique for biomarker discovery and exposomics research and demonstrates the capability of the Breath Biopsy Platform to collect breath samples for the discovery of VOC biomarkers relevant to the exposome.