Validating differential volatilome profiles in Parkinson's disease

Eleanor Sinclair[1], Caitlin Walton-Doyle[1], Depanjan Sarkar[1], Katherine A Hollywood[1,2], Joy Milne[1], Sze Hway Lim[3], Tilo Kunath[4], Anouk M. Rijs[5], Rob M. A. de Bie[6], Monty Silverdale[3], Drupad K Trivedi[1] and Perdita Barran[1]

[1]Manchester Institute of Biotechnology, School of Chemistry, The University of Manchester, Princess Street, Manchester, UK, M1 7DN
[2]Manchester Centre for Synthetic Biology of Fine and Speciality Chemicals (SYNBIOCHEM), Manchester Institute of Biotechnology, School of Chemistry, The University of Manchester, Princess Street, Manchester, UK, M1 7DN
[3]Department of Neurology, Salford Royal Foundation Trust, Manchester Academic Health Science Centre, University of Manchester, UK
[4]Institute for Stem Cell Research, School of Biological Sciences, The University of Edinburgh, Edinburgh UK, EH16 4UU
[5]Radboud University, Institute for Molecules and Materials, FELIX Laboratory, Toernooiveld 7c, 6525 ED, Nijmegen, The Netherlands
[6]Amsterdam University Medical Centers, University of Amsterdam, Department of Neurology, Amsterdam Neuroscience, Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands

Abstract

Parkinson's disease (PD) is a progressive neurodegenerative disorder that does not currently have a robust clinical diagnostic test. Non-motor symptoms such as skin disorders have long since been associated with the disease and more recently a characteristic odour emanating from the skin of people with Parkinson's has been identified. Here, dynamic head space (DHS) thermal desorption (TD) gas chromatography-mass spectrometry (GC-MS is implemented to directly measure the volatile components of sebum on swabs sampled from people with Parkinson's – both drug naïve and those on PD medications (*n*= 100) and control subjects (*n*= 29). Supervised multivariate analyses of data showed 84.4 % correct classification of PD cases using all detected volatile compounds. Variable importance in projection (VIP) scores were generated from these data, which revealed eight features with VIP > 1 and $p < 0.05$ which all presented a downregulation within the control cohorts. Purified standards based on previously annotated biomarkers eicosane and octadecanal did not match to clinical data, although multiple metabolite features are annotated with these compounds all with high spectral matches indicating the presence of a series of similar structured species. DHS-

TD-GC-MS analysis of a range of lipid standards has revealed the presence of common hydrocarbon species rather than differentiated intact compounds which are hypothesised to be breakdown products of lipids. This replication study validates that a differential volatile profile between control and PD cohorts can be measured using an analytical method that measures volatile compounds directly from skin swabs.

Introduction

The use of scent as an indicator for physical (and mental) biological functions has been associated with many disease states, such as diabetes *mellitus*, tuberculosis, liver and kidney disease and cancer.[1] Although odour can be linked to many diseases it is not a commonly applied approach to define biomarkers in clinical diagnostic tests. The application of canine detection within the medical field has progressed since a pioneering report of Melanoma detection by Williams and Pembroke in 1989.[2–4] Some characteristic odours released in bodily secretions and excretions can occur before many other symptoms of the disease have developed as previously reported in Parkinson's disease (PD).[5,6]

Odorous skin secretions are most frequently related to areas of high sweat excretion and its subsequent bacterial degradation. However, sebaceous gland excretion is an alternative process that produces a biological matrix with a signature due to volatile organic compound(s) (VOC(s)).[7] The production of sebum from these glands occurs everywhere on the body, with the exception of the palms of hands and soles of feet and is most evident on the face and upper trunk of the body which are categorised as sebum rich locations.[8]

A noteworthy non-motor symptom of PD is the development of skin related disorders such as seborrheic dermatitis (SD), which occurs in up to 60 % of PD patients and manifests as an increase in sebum production.[9,10] The analysis of volatile species from skin secretions such as sebum and sweat is an underdeveloped methodology for clinical applications and hence there is no recognised standard procedure for its sampling.[11,12] Gauze swabs,[13] polydimethylsiloxane (PDMS) patches,[14] cigarette papers[15], glass rollers or beads[16,17] and more advanced methods such as wearable iontophoresis biosensors[18] have been previously reported, all of which have associated advantages and disadvantages in application.[19] These

methods have primarily been directed to the collection of sweat rather than sebum, the waxy lipid-rich component.

The chemicals associated with these odorous profiles belong to the VOC class. There has been a rise in the development of analytical techniques to measure VOCs for clinical applications, most notably in the advancement of 'electronic noses' and within the field of breathomics.[20–24] Analysis of VOCs using gas chromatography-mass spectrometry (GC-MS) is a popular non-selective analytical technique in which complex mixtures are separated prior to *m/z* detection of molecular and corresponding fragment ions to aid in structural elucidation. Headspace analysis, either dynamic or static, is a technique which promotes the volatile components of a sample to enter the gaseous 'headspace' of an enclosed vessel by the partition of a concentration gradient, often using a temperature incubation prior to sampling to enhance the process.[25,26] Whilst both techniques are effective, dynamic headspace (DHS) analysis has the benefit of concentrating volatile species through the continuous collection of the headspace which is an obvious benefit for untargeted metabolomics.[25]

This study leads on from earlier work which demonstrated that there was a differential VOC signature in Parkinson's disease patients.[5] This subsequent study provides a completely independent validation set of data using a different instrument alongside new patient samples.

Experimental Methods

Sample Participants: The participants included within this study were part of a nationwide recruitment process taking place at 25 different NHS clinics. This study consists of a subset of 129 participants from three subject groups; independent controls (*n*=29), drug naïve Parkinson's disease participants (*n*=17) and medicated Parkinson's disease participants (*n*=83). Patient demographics are reported in Table S1. Ethical approval for this project (IRAS project ID 191917) was obtained by the NHS Research Authority (REC references: 15/SW/0354).

Sample Collection: Each participant was swabbed on the upper back with cotton-based sterile medical gauze (7.5 cm x 7.5 cm) to collect sebum present on the skin. Participants were asked

not to wash 24 hrs before sebum collection. The patient sampled gauze was sealed in background-inert plastic bags and transported to the central facility at the University of Manchester, where they were stored at -80 ^{o}C until analysis.

Chemicals and Materials: The chemicals and materials used in this study were: gauze swabs (Arco, UK), sterile sample bags (GE Healthcare Whatman^{TM}, UK), 20 mL glass vials (GERSTEL, Germany), TENAX TA thermal desorption tubes and liners packed with TENAX TA for CIS 4/6 liner for TDU (GERSTEL, Germany), Optima® LC-MS grade methanol (Fisher Scientific), HiPerSolv CHROMANORM® absolute ethanol 99.8 % purity (VWR Chemicals). The QC sample was comprised of a mixture of seven compounds, each sourced from Sigma Aldrich; L(-)-carvone (27.0 µM), δ-decalactone (96.4 µM), ethyl butyrate (150.5 µM), ethyl hexanoate (30.5 µM), hexadecane (43.7 µM), nonane (83.07 µM), vanillin (100.8 µM), all in MeOH:EtOH (9:1). Octadecanal (Apollo Scientific) and eicosane (Sigma Aldrich) standards were both 99 % purity.

DHS-TD-GC-MS Analytical Method: Thawed gauze swabs were aseptically transferred into 20 mL glass vials and analysed by DHS-TD-GC-MS. During DHS pre-concentration the samples were incubated at 80 °C for 10 minutes to promote concentration of VOCs in the vial headspace. Trapping was initiated with dry nitrogen as the purge gas at a flow rate of 70 mL.min^{-1} for a total gas volume of 1 L. The volatile compounds were captured on a TENAX TA adsorbent tube (GERSTEL, Germany) held at 40 °C. The adsorbent tube was transferred from the DHS unit to the TD unit (TDU) using an automated GERSTEL MPS dual head workstation, at which time the captured VOCs were desorbed from the TENAX sorbent. The TDU was operated in splitless mode and was held at 30 °C for one min before the application of a 12 °C.s^{-1} temperature ramp and held at 280 °C for five min. The desorbed analytes were focused in a cooled injection system (CIS) which was operated in solvent vent mode, using a vent flow of 80 mL.min^{-1}. The CIS was held at 10 °C for two minutes before a second 12 °C.s^{-1} temperature ramp and maintained at 280 °C for five min. The CIS vent valve (and flow) was not initiated until three min after analyte release on the GC column and can therefore be classed as a splitless method. The GC analysis was performed on an Agilent GC 7890A coupled to an Agilent MSD 5975 interfaced by an electron impact (EI) source. Separation was induced by an Agilent VF-5MS column (30 m x 0.25 mm x 0.25 µm). The column flow was kept at 1 mL.min^{-1}. The oven ramp was programmed as follows: 40 °C held for one min, 25 °C.min^{-1} to

180 °C, 8 °C.min^{-1} to 240 °C and held for one min, 20 °C.min^{-1} to 300 °C held for 2.9 min for a total run time of 21 min. The transfer line to the MS was maintained at 300 °C, the EI source at 230°C and the Quadrupole at 150°C. The MSD operated in scan mode for a mass range between 30 and 800 *m/z*. The GC-MS system is fitted with a GERSTEL olfactory detection port (ODP3) using Agilent Technologies capillary flow technology (three-way splitter plate equipped with makeup gas). This was controlled as a 1:1 (*v:v*) split between MSD and ODP and therefore the GC eluent was diluted by a factor of two with helium makeup gas.

Clinical samples analysis: Gauze samples from 129 subjects were analysed across seven stratified, randomised and blinded analytical batches. Quality Control (QC) samples were injected at the beginning (*n*=3), after every 5th injection and at the end (*n*=3) of each batch analysis. QC samples (5 µL) were used to gauge the analytical reproducibility in batch-to-batch analyses, which were analysed on a shorter DHS-TD-GC-MS method to enable a faster analysis time per analysis batch.

Candidate biomarker standards analysis: Octadecanal and eicosane standards were run on an identical DHS-TD-GC-MS method to clinical samples. Headspace sampling was performed with 10 µL of octadecanal (37.4 µM) and eicosane (35.4 µM) in MeOH and were individually analysed in 20 mL vials.

Lipid standards analysis: Lipid standards were purchased from Avanti Polar Lipids; L-α-phosphatidylethanolamine (Brain, Porcine) (PE), L-α-phosphatidylserine (brain, porcine) (sodium salt) (PS) and 1',3'-bis[1,2-dioleoyl-sn-glycero-3-phospho]-glycerol (sodium salt) (18:1 cardiolipin) (CL) and were diluted in $CHCl_3$ whilst L-α-phosphatidylcholine (brain, porcine) (PC) was diluted in MeOH. Glucosylsphingosine (GlcSph) (Sigma Aldrich) and was diluted in MeOH. All standards were analysed using a slightly varied analytical method to that of the clinical samples, the changes were as follows: a five-minute solvent vent was applied in addition to a further four-minute-high temperature hold at the end of the GC temperature gradient. The analytical method was 25 minutes in total. From a standard solution of 100 µM total, 20µL were analysed in separate headspace vials. Lipid standards were measured in triplicate alongside blank headspace vials and blank solvent analyses (both MeOH and $CHCl_3$) which were measured both prior to and post lipid standard analysis.

Data Pre-Processing and Deconvolution: All TD−GC−MS clinical and lipid standards data were converted to open source mzML format using ProteoWizard.[27] The dataset was deconvolved using in-house script using eRah package for R.[28] The deconvolved analytes were assigned putative identifications by matching fragment spectra with compound spectra using the Golm database. The resulting matrices comprised of variables and their corresponding peak areas per sample. The clinical dataset comprised 671 features and these were further refined by the removal of features absent in more that 5 % of all samples, which generated a reduced dataset of 520 features.

Features present in lipid standards data were filtered based upon a set of further criteria: (a) match factor (MF) > 75, (b) detection only in lipid sample analysis and not present in either blank headspace vial or solvent-only analyses, (c) abundance > 1000 counts and (d) each feature was present in at least two of three triplicate analyses.

Statistical Analysis: We used a two-pronged approach with statistical investigation of data. A data-driven multivariate approach was used to validate if there was a differential VOC signature associated with PD. Further, a targeted approach was used to verify selective biomarker peaks. Data were auto-scaled and all missing values replaced using spline interpolation prior to statistical analysis. Synthetic minority over-sampling technique (SMOTE) was implemented to balance the sample numbers within each class and hence remove possible bias to the majority class.[29] Partial least squares-discriminant analysis (PLS-DA) was executed in MATLAB (2019a) for classification and prediction of classes within data.[30,31] Models were validated by resampling bootstrapping (*n*=250). Multivariate Receiver Operating Characteristic (ROC) analysis was performed using MetaboAnalyst Biomarker Analysis (Version 4.0)[32], PLS-DA was implemented for classification and feature ranking with a two latent variable input. ROC curves were generated by balanced Monte Carlo cross validations (MCCV) in which two thirds of the samples were used to evaluate feature performance and the remaining one third were used to validate the classification. Iterations of this process (*n*=30) were performed to calculate model performance and calculate confidence intervals for area under the curve (AUC) for each model.[33]

Results and Discussion

Validating Changes within VOC Profiles with the Onset of PD

To assess and validate if there are measurable discriminant VOCs in sebum, PLS-DA models were generated from new independent data. Models were generated to analyse (i) the classification accuracy between PD (drug naïve and medicated) and control samples and (ii) drug naïve PD against medicated PD participants to investigate differences in the disease phenotype alongside any possible effects of medication on this measured phenotype. An average correct classification rate (CCR) of 84.4 % was obtained by combining the drug naïve and medicated PD cohorts in which the CCR was validated *via* bootstrapping (*n*=250) (Figure 1). This prediction rate is similar to the average CCR (86 %) achieved by PLS-DA with 5-fold cross validation reported in our pilot study.[5] Patient demographics were assessed to determine if they impacted the classification accuracy within the model. This is discussed in detail within the Supplementary Information alongside significance tests for metadata parameters (Table S2). In summary, no significant confounding effects were observed on PLS-DA classification due to gender (Figure S1), BMI, alcohol intake or smoking (Figure S2) or age.

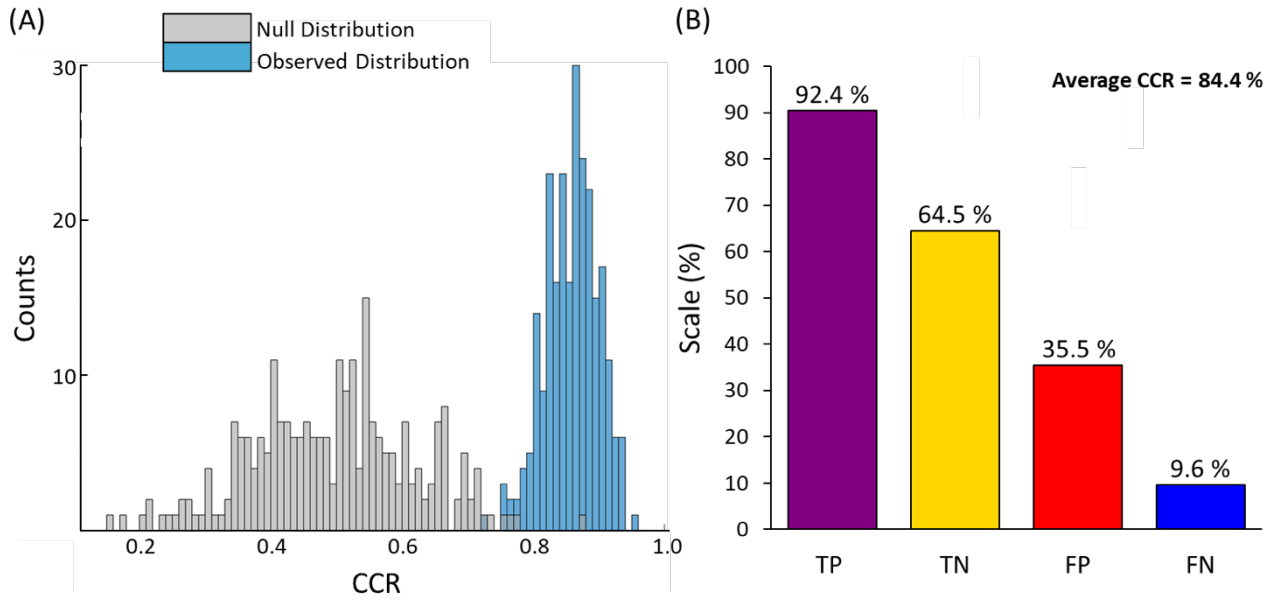

Figure 1: PLS-DA classification model for a two-class input using combined drug naïve and medicated PD cohorts *vs.* controls. (A) Histogram reporting the distribution of the correct classification rate for the null (grey) and observed (blue) distributions obtained from

bootstrap validation (*n*=250) of the PLS-DA classification model. (B) A chart displaying the true positive (TP), true negative (TN), false positive (FP) and false negative (FN) classification rates.

This classification validates findings from our initial TD-GC-MS study which revealed a unique VOC signature that can be associated to PD patients and be measured by the headspace sampling of sebum collected non-invasively. It is noted that the measured VOC profiles between drug naïve PD and medicated PD cohorts did not classify well, with a CCR of 66.7 % and had poor validation by permutation testing (Figure 2). This indicates that the sampled VOCs do not have sufficient discriminatory power to classify between drug naïve and medicated PD but can distinguish PD from control, indicating that the associated odour profile is fundamental to PD. It can therefore be hypothesised that there is minimal change within the VOC profile after the onset of PD, with treatment, and/or that our classification of the disease based on drug naïve *vs.* medication does not translate well to clinical PD staging.

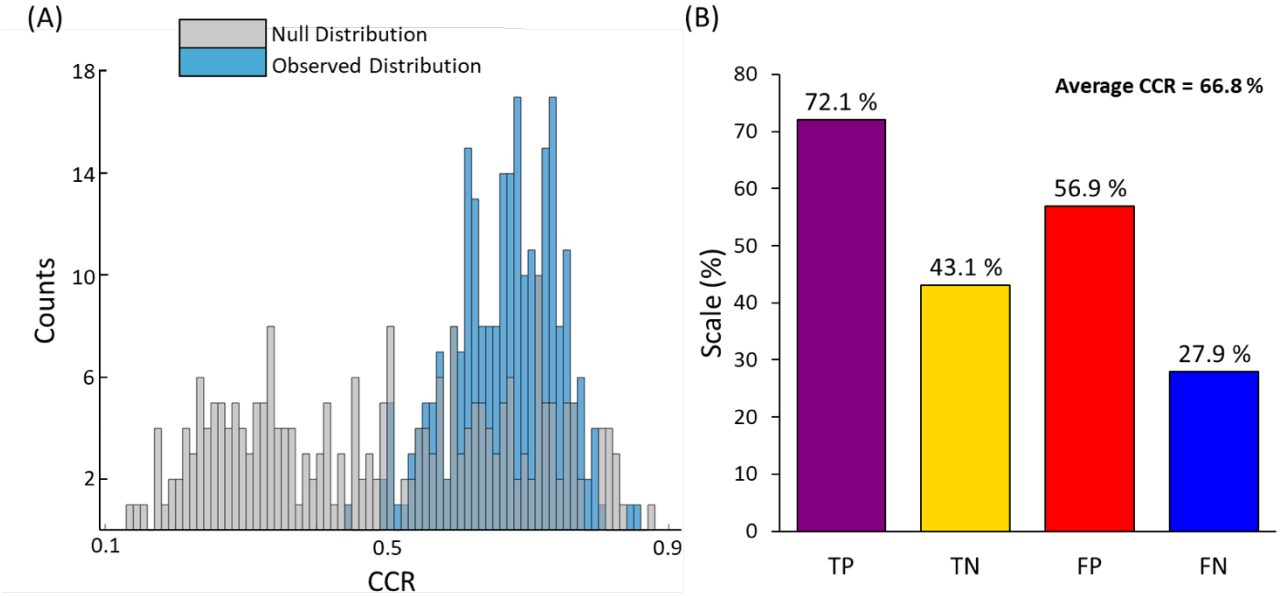

Figure 2: PLS-DA classification of drug naïve PD *vs.* medicated PD in which medicated PD was the positive predictive class (A) Histogram reporting the distribution of Correct Classification Rates (CCR) for the null (grey) and observed (blue) distributions obtained from bootstrap validation (*n*=250) of the PLS-DA classification model. (B) A chart reporting the classification rates from the PLS-DA model.

Discriminatory Features in PLS-DA Classification Modelling

Variable importance in projection (VIP) scores were calculated for the two-class PLS-DA model: combined PD *vs.* controls. A multivariate ROC analysis was performed for compounds (*n*=12) with VIP score > 1 (Figure 3A). The area under the curve (AUC) values increase as the number of features included in the model increases, in conjunction with a reduction in the confidence interval (CI) range. The predictive power of model increases going from a two-variable classifier up to seven-variable classifier, however, begins to plateau following this threshold. The compounds that had high VIP scores (top 10 %) were further selected (*n*=50) and ROC analysis was performed, showing the same trend of increasing AUC values until a plateau between the five-variable and ten-variable models, which is consistent with the previous ROC analysis (VIP > 1). This is presented in Supplementary Figure S3.

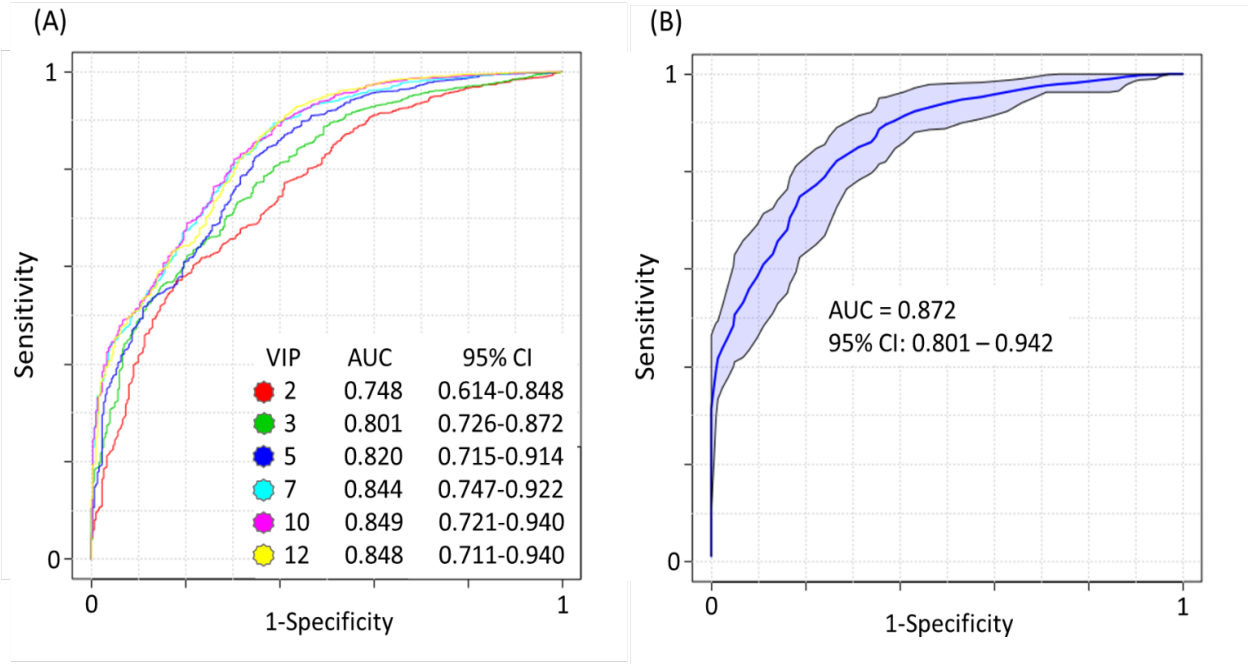

Figure 3: Multivariate ROC curve analyses to evaluate the performance of VIP compounds in biomarker models. Resampling was used to calculate 95 % confidence intervals (CI) using a Monte Carlo Cross Validation (MCCV) approach. (A) Each coloured line represents the ROC curve using a specific number of variables, these are listed in the bottom right hand corner from red (two-variable) to yellow (12-variable), generated from all compounds with VIP > 1. (B) ROC curve displaying the sensitivity and specificity of all eight-variables (VIP >1, p <0.05) combined; the shaded bands represent the 95 % CI for this model.

Mann-Whitney non-parametric U tests between PD and control cohorts were performed for selected compounds (VIP > 1) and eight of these compounds were found to be statistically significant with $p < 0.05$. ROC analysis of these compounds provides an improved classification model (Figure 3B) in which the addition of all eight features improves the AUC to 0.872 and the lowest CI range (0.801-0.942). Figure S4 displays the individual ROC curve for each added compound in this model. Due to limitations of volatile compound identification within this study, species are simply referred to as 'VIP features' (1-8). Box and whiskers plots for these eight compounds are reported in Figure S5 alongside calculated fold changes (non-scaled). Each feature is downregulated within the PD cohort with fold changes ranging from 0.2 to 0.73 between PD and control cohorts.

Targeted Analysis of Previously Identified Biomarkers

The following stage of the study concerned investigating the presence and significance of a panel of candidate biomarkers obtained from the results of our previous study.[34] A targeted data analysis approach was performed for the four target compounds: octadecanal, eicosane, hippuric acid and perillaldehyde. The data matrix revealed that multiple features were annotated with the same compound identification for each of these targets based upon high similarities of their fragmentation pattern signatures. Database spectral matching typically returns a match factor (MF), in this case between 1 and 100, which indicates the similarity of an experimentally recorded spectrum to that of a database spectrum based on the presence of molecular and fragment ion peaks and their relative intensities. Eicosane and octadecanal each returned three identification matches, respectively, in which these annotations were the primary ID, falling within a two to three-minute retention time (RT) window. Standards of eicosane and octadecanal were analysed using an identical analytical method and the corresponding reference spectra were matched against these highlighted features within the clinical sebum data. Although retention times could not be matched, there was a high similarity between mass spectra (MF >90). It can be inferred that these putative assignments based on spectra matching result from closely related molecular species that have large hydrocarbon chains, for example lipids. Sebum is composed of large number of lipid species such as triglycerides, fatty acids and wax esters.[35,36] Thus, it is not surprising to observe multiple closely related lipid peaks.

To corroborate this inference, lipid standards were purchased and analysed to investigate the resultant chromatographic spectra from known standards. Evidence from two parallel studies analysing PD and control sebum, using liquid chromatography-MS[37] and paper spray-MS,[38] have shown the primary compounds of interest can be putatively assigned to a series of lipids. Lipid mixtures were investigated to distinguish the type of species detected in (non-derivatised) lipid DHS-TD-GC-MS experiments and to then evaluate their compound annotation in the same libraries. Each lipid standard mixture yielded a large number of resolvable compound features across each chromatogram that had high database spectral MF scores and good reproducibility across replicates. Between 68 % and 87 % of these features were annotated as hydrocarbon species assigned to at least two detected variables. A list of the hydrocarbon species detected within the lipids standards data is reported in Table 1, alongside the number of discrete features provided with this annotation. Definitive annotation of large, chemically similar hydrocarbons, like lipids is challenging because of the obvious similarities expected in their fragment products. This is further confounded by the large number of conceivable structural isomers for each hydrocarbon, for example, dodecane the most frequently detected species in the data has 355 possible isomers. These experiments were repeated using a traditional liquid injection GC-MS experimental setup for the analysis of some of these lipid standards (Table S3) and a similar trend in recurring hydrocarbon features was noted. We hypothesise that the overlap of compound annotations across distinct deconvolved features in our clinical sebum data arises from the decomposition of lipid species.

For example, lipid oxidation is a well-known mechanism for lipid degradation and can proceed via an assortment of pathways dependent on the lipid and enzyme in question and the type of oxidation. These reactions produce a variety of product species including hydrocarbons, alongside aldehydes, ketones, alcohols, esters and acids.[39,40] Therefore, the multiple instances of hydrocarbon species that are reproducibly detected throughout lipid-based chromatographic analyses (both clinical and from standards) are due to lipid decomposition. Such degradation can lead to an accumulation of large concentrations of common hydrocarbon chains of different lengths. Similar products due to the loss of more distinctive structural moieties, such a lipid head groups, are readily observed upon collisional activation and in source dissociation of these labile compounds. These results provide a vital

insight into the expected profile upon analysis of the hydrocarbon rich lipid-based biological matrix that is sebum. The identification of pure, large molecular weight lipids standards by GC-MS is challenging and data shown here yields an array of product species across the chromatographic domain, unquestionably the myriad of endogenous and exogenous compounds in sebum will offer a greater challenge. Prior to the generation of comprehensive annotated databases that reflect this chemical complexity, it is likely that biomarker detection will consist of Metabolomics Standards Initiative (MSI) level 2 feature detection rather than the verification of compounds.

Table 1: Putative ID and chemical formula of the hydrocarbon species present within the lipid standards data. The number of distinct features that are annotated with this compound ID are listed beside.

		Number of features				
Putative ID	**Chemical Formula**	**PE**	**PC**	**PS**	**CL**	**GlcSph**
Decane	C10H22	2	0	2	2	0
Dodecane	C12H26	9	4	6	7	6
Tridecane	C13H28	1	0	0	1	0
Tetradecane	C14H30	2	0	2	0	0
Pentadecane	C15H32	2	0	2	2	2
Hexadecane	C16H34	6	1	2	6	1
Heptadecane	C17H36	2	2	4	3	1
Nonadecane	C19H40	0	0	1	2	0
Eicosane	C20H42	2	2	4	3	0
Heneicosane	C21H44	0	0	0	0	1
Docosane	C22H46	0	1	0	1	0
Tricosane	C23H48	3	0	3	2	0
Tetracosane	C24H50	0	0	0	0	2

The lipid abbreviations are as follows; PE: Phosphatidylethanolamine; PC: Phosphatidylcholine; PS: Phosphatidylserine; CL: Cardiolipin (18:1) and GlcSph: Glucosylsphingosine

The remaining candidate biomarker compounds - hippuric acid and perillaldehyde - could only be identified in a derivatised form with trimethylsilyl and methoxime adducts, respectively. Again, multiple features were assigned these annotations and they spanned a wide chromatographic range, therefore annotation of neither of these prospective biomarkers could be reliably verified within this dataset. A direct limitation of analysing a biological matrix without derivatisation is the limited compound annotation using commercial databases, which comprise majorly of derivatised compounds. The dominating method in GC-MS analysis for metabolomics approaches uses chemically derivatised sample extracts and

therefore biologically relevant compound spectra are often of derivatised forms. Further work is needed in the area of database curation for non-derivatised, biologically relevant, volatile species present in a sebum as a biological matrix, to enable confident compound identification using DHS-GC-TD-MS.

Conclusion

Volatile organic compounds from sebum measured using DHS-TD-GC-MS can accurately discriminate between PD and control samples. Our results validate previous findings that VOCs measured from skin have a differential profile in Parkinson's disease that can be modelled using PLS-DA analysis. Putatively annotated biomarker candidates from our previous study could not be validated within this dataset. However, multiple features were putatively annotated as two of these candidates (octadecanal and eicosane) and derivatised compound annotations were associated to the remaining two biomarkers (hippuric acid and perillaldehyde). Spectral matching to an in-house database of standards *viz.* octadecanal and eicosane yielded high spectral similarity (MF > 90), although retention times did not match. This leads us to conclude that these compounds have very similar structures and/or could be common breakdown products of larger species in which the discriminatory structural moiety was lost as a neutral fragment. This hypothesis was strengthened with the analysis of lipid standard mixtures, which reveal the detection of common hydrocarbon species. Compounds selected using VIP scores-based ranking from PLS-DA modelling, performed well in multivariate ROC analysis and each metabolite feature showed a fold change relating to a lower expression in PD samples. These compounds have not been annotated with putative identifications due to poor performance of database matching to available GC-MS mass spectra libraries. Future studies will create an in-house GC-MS database for both spectral and RT matching that will address current bottleneck in the field for annotation of non-derivatised, biologically relevant VOCs.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgements

We thank Michael J Fox Foundation (grant ref:12921) and Parkinson's UK (grant ref: K-1504) for funding this study. This work was supported by an EPSRC DTA grant to the School of Chemistry, which has funded the PhD project of ES and the BBSRC (award BB/L015048/1) for instrumentation used in this work. We also thank our recruitment centres for their enthusiasm and rigor during the recruitment process. We are very grateful to all the participants who took part in this study as well as PIs and nurses across all the recruiting centres. We also thank Richard Weller for feedback and discussions on sebum and dermatology.

References

1. Shirasu, M. & Touhara, K. The scent of disease : volatile organic compounds of the human body related to disease and disorder. *J. Biochem.* **150**, 257–266 (2011).
2. Williams, H. & Pembroke, A. Sniffer dogs in the Melanoma Clinic? *Lancet* **9285**, 734 (1989).
3. Willis, C. M. *et al.* Olfactory Detection of Human Bladder Cancer by Dogs: Proof of Principle Study. *BMJ* **329**, 1–6 (2004).
4. Pirrone, F. & Albertini, M. Olfactory detection of cancer by trained sniffer dogs: A systematic review of the literature. *J. Vet. Behav. Clin. Appl. Res.* **19**, 105–117 (2017).
5. Trivedi, D. K. *et al.* Discovery of Volatile Biomarkers of Parkinson's Disease from Sebum. *ACS Cent. Sci.* **5**, 599–606 (2019).
6. Morgan, J. Joy of super smeller: sebum clues for PD diagnostics. *Lancet Neurol.* **15**, 138–139 (2016).
7. Gallagher, M. *et al.* Analyses of volatile organic compounds from human skin. *Br. J. Dermatol.* **159**, 780–791 (2008).
8. Borda, L. J. & Wikramanayake, T. C. Seborrheic Dermatitis and Dandruff: A Comprehensive Review. *J. Clin. Investig. Dermatology* **3**, 1–22 (2015).
9. Arsic Arsenijevic, V. S. *et al.* A laboratory-based study on patients with Parkinson's disease and seborrheic dermatitis: the presence and density of Malassezia yeasts, their different species and enzymes production. *BMC Dermatol.* **14**, 5 (2014).
10. Krestin, D. The seborrhoeic facies as a manifestation of post-encephalitic parkinsonism and allied disorders. *QJM* **21**, 177–186 (1927).
11. Jadoon, S. *et al.* Recent developments in sweat analysis and its applications. *Int. J. Anal. Chem.* **164974**, (2015).
12. Hussain, J. N., Mantri, N. & Cohen, M. M. Working up a good sweat - The challenges

of standardising sweat collection for metabolomics analysis. *Clin. Biochem. Rev.* **38**, 13–34 (2017).

13. Curran, A. M., Rabin, S. I., Prada, P. A. & Furton, K. G. Comparison of the volatile organic compounds present in human odor using SPME-GC/MS. *J. Chem. Ecol.* **31**, 1607–1619 (2005).

14. Thomas, A. N. *et al.* Novel noninvasive identification of biomarkers by analytical profiling of chronic wounds using volatile organic compounds. *Wound Repair Regen.* **18**, 391–400 (2010).

15. Shetage, S. S. *et al.* Effect of ethnicity, gender and age on the amount and composition of residual skin surface components derived from sebum, sweat and epidermal lipids. *Ski. Res. Technol.* **20**, 97–107 (2014).

16. Bernier, U. R., Kline, D. L., Barnard, D. R., Schreck, C. E. & Yost, R. A. Analysis of human skin emanations by gas chromatography/mass spectrometry. 2. Identification of volatile compounds that are candidate attractants for the yellow fever mosquito (Aedes aegypti). *Anal. Chem.* **72**, 747–756 (2000).

17. Kutyshenko, V. P., Molchanov, M., Beskaravayny, P., Uversky, V. N. & Timchenko, M. A. Analyzing and mapping sweat metabolomics by high-resolution NMR spectroscopy. *PLoS One* **6**, 4–12 (2011).

18. Emaminejad, S. *et al.* Autonomous sweat extraction and analysis applied to cystic fibrosis and glucose monitoring using a fully integrated wearable platform. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 4625–4630 (2017).

19. Riazanskaia, S., Blackburn, G., Harker, M., Taylor, D. & Thomas, C. L. P. The analytical utility of thermally desorbed polydimethylsilicone membranes for in-vivo sampling of volatile organic compounds in and on human skin. *Analyst* **133**, 1020–1027 (2008).

20. Turner, A. P. & Magan, N. Electronic noses and disease diagnostics. *Nat. Rev. Microbiol.* **2**, 161–166 (2004).

21. Bach, J. P. *et al.* Measuring compounds in exhaled air to detect Alzheimer's disease and Parkinson's disease. *PLoS One* **10**, 1–13 (2015).

22. D'amico, A. *et al.* Identification of melanoma with a gas sensor array. *Ski. Res. Technol.* **14**, 226–236 (2008).

23. Rattray, N. J. W., Hamrang, Z., Trivedi, D. K., Goodacre, R. & Fowler, S. J. Taking your breath away: Metabolomics breathes life in to personalized medicine. *Trends Biotechnol.* **32**, 538–548 (2014).

24. Beale, D. J. *et al.* A review of analytical techniques and their application in disease diagnosis in breathomics and salivaomics research. *Int. J. Mol. Sci.* **18**, 1–26 (2017).

25. Wang, Y., McCaffrey, J. & Norwood, D. L. Recent advances in headspace gas chromatography. *J. Liq. Chromatogr. Relat. Technol.* **31**, 1823–1851 (2008).

26. Sithersingh, M. J. & Snow, N. H. Headspace-Gas Chromatography. in *Gas Chromatography* (ed. Poole, C.) 221–233 (Elsevier Inc., 2012). doi:10.1016/b978-0-12-

385540-4.00030-4

27. Chambers, M. C. *et al.* A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **30**, 918–920 (2012).

28. Domingo-Almenara, X. *et al.* eRah: A Computational Tool Integrating Spectral Deconvolution and Alignment with Quantification and Identification of Metabolites in GC/MS-Based Metabolomics. *Anal. Chem.* **88**, 9821–9829 (2016).

29. Chawla, N. V, Bowyer, K. W., Hall, L. O. & Kegelmeyer, P. W. SMOTE: Synthetic Minority Over-Sampling Technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002).

30. Xu, Y., Muhamadali, H., Sayqal, A., Dixon, N. & Goodacre, R. Partial least squares with structured output for modelling the metabolomics data obtained from complex experimental designs: A study into the ϒ-block coding. *Metabolites* **6**, 38 (2016).

31. Xu, Y. & Goodacre, R. On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. *J. Anal. Test.* **2**, 249–262 (2018).

32. Chong, J. *et al.* MetaboAnalyst 4.0: Towards more transparent and integrative metabolomics analysis. *Nucleic Acids Res.* **46**, W486–W494 (2018).

33. Xia, J., Broadhurst, D. I., Wilson, M. & Wishart, D. S. Translational biomarker discovery in clinical metabolomics: An introductory tutorial. *Metabolomics* **9**, 280–299 (2013).

34. Trivedi, D. K. *et al.* Discovery of Volatile Biomarkers of Parkinson's Disease from Sebum. *ACS Cent. Sci.* **5**, 599–606 (2019).

35. Picardo, M. & et al. Sebaceous gland lipids. *Dermato-Endocronology* **1**, 68–71 (2009).

36. Pappas, A. Epidermal surface lipids. *Dermato-Endocronology* **1**, 72–76 (2009).

37. Sinclair, E. *et al.* Sebum: A Window into Dysregulation of Mitochondrial Metabolism in Parkinson's Disease. *ChemRxiv* (2020). doi:10.26434/chemrxiv.11603613

38. Sarkar, D. *et al.* Paper Spray of Sebum : A Step Towards Rapid Confirmatory Diagnosis of Parkinson's Disease. Manuscript in Preparation (2020).

39. Ross, C. F. & Smith, D. M. Use of volatiles as indicators of lipid oxidation in muscle foods. *Compr. Rev. Food Sci. Food Saf.* **5**, 18–25 (2006).

40. Domínguez, R. *et al.* A comprehensive review on lipid oxidation in meat and meat products. *Antioxidants* **8**, 1–31 (2019).