Discovery of volatile biomarkers of Parkinson's disease from sebum

Drupad K Trivedi[1], Eleanor Sinclair[1], Yun Xu[1], Depanjan Sarkar[1], Camilla Liscio[2], Phine Banks[2], Joy Milne[1], Monty Silverdale[3], Tilo Kunath[4], Royston Goodacre[1,5], Perdita Barran*[1]

[1]Manchester Institute of Biotechnology, School of Chemistry, The University of Manchester, Princess Street, Manchester, UK, M1 7DN
[2]Anatune, 4 Wellbrook Way, Girton, Cambridge, UK, CB3 0NA
[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] Institute of Integrative Biology, University of Liverpool, Crown Street, Liverpool L69 7ZB

Supplementary Information

Table S1A Details of the collecting sites in the UK and the lead PI at each site

SITE NUMBER	SITE NAME	PI
1	Addenbrookes (Cambridge)	Paul Worth
2	Bournemouth	Khaled Amar
3	Cornwall/Truro	Christine Schofield
4	Lothian - Western General Edinburgh	Gordon Duncan
5	Edinburgh – Royal Infirmary of Edinburgh	Gordon Duncan
6	Edinburgh - Primary Care NHS Lothian (Seb Derm)	Richard Weller
7	Hampshire	Sam Arianayagam
8	Nottingham	Gill Sare
9	Pennine	Jason Raw
10	Salford	Monty Silverdale
11	Salisbury	Diran Padiachy
12	Sheffield	Oliver Bandmann
13	South Tees	Neil Archibold
14	Southern Health	Helen Roberts
15	Luton & Dunstable	Anette Schrag
16	Portsmouth	Sean Slaght
17	Northumbria	Richard Walker

18	London North West	Sophie Molloy
19	Bath	Veronica Lyell
20	Gateshead	Richard Athey
21	Sunderland	Uma Nath
22	Plymouth	Camille Caroll
23	Newcastle Upon Tyne Hospitals NHS Foundation Trust (Newcastle University)	Nicola Pavese
24	Royal Devon and Exeter NHS Foundation Trust	Robert James
25	Imperial College Healthcare NHS Trust	Sophie Molloy

Table S1B: Participant numbers and metadata per wave.

		Discovery cohort		
	Control (*n*=10)	Drug Naïve PD (*n*=10)	PD on medication (*n*=10)	*p*-value
Age (years)	64.8 ± 3.06	72.82 ± 8.42	64.67 ± 2.55	0.01*
BMI	27.10 ± 3.50	26.94 ± 4.08	25.33 ± 3.44	0.64
Gender (M/F ratio)	0.84	1.20	0.80	0.88
Alcohol intake (yes/no ratio)	4.5	0.37	2	0.03*
Smoker	1	0	0	0.39
		Validation cohort		
	Control (*n*=11)	Drug Naïve PD (*n*=11)	PD on medication (*n*=9)	*p*-value
Age (years)	55.78 ± 18.87	75.40 ± 6.85	68.90 ± 11.76	0.02*
BMI	28.96 ± 11.01	25.74 ± 3.83	24.98 ± 3.54	1.00
Gender (M/F ratio)	0.26	1.50	1	0.10
Alcohol intake (yes/no ratio)	0.8	9	1.5	0.10
Smoker	0	0	1	0.24

Wave 3 (odor port validation, drug naïve PD subjects only, *n*=3)	
Age (years)	65.66 ± 3.30
BMI	23.46 ± 1.80
Gender (M/F ratio)	2
Alcohol intake (yes/no ratio)	2
Smoker	0

* indicates significant difference between controls, drug naïve and PD with medication groups.

Table S2: List of candidate volatiles putatively identified (MSI level 2) and matched across two different cohorts. Nine of out 17 metabolites listed were selected for further analysis since they had acceptable retention time drift between the two sets of experiments.

Putative identification	Mass	Retention time (min) (discovery)	Retention time (min) (validation)	Retention time (min) difference	Comments
3,4-dihydroxy mandelic acid	184.15	20.87	Not found	n/a	Not found
Artemisinic acid	234.34	12.97	12.83	0.14	Included
Cyclohexasiloxane, dodecamethyl	357.57	16.47	16.06	0.41	Excluded
Cyclohexylcyclohexane	357.57	15.36	14.71	0.65	Excluded
Dodecane	170.34	13.20	13.27	-0.07	Included
Eicosane	282.56	20.65	20.62	0.03	Included
Gallic acid ethyl ester	198.17	11.40	10.99	0.41	Excluded
Glutamine	128.09	21.73	21.09	0.64	Excluded
Hexyl acetate	170.34	11.70	11.53	0.16	Included
Hippuric acid	179.17	20.61	20.52	0.09	Included
Neoabietic acid	302.46	21.66	Not found	n/a	Not found
Octacosane	394.77	17.49	17.46	0.03	Included
Octadecanal	170.34	20.87	20.75	0.12	Included
Octanal	244.38	11.58	11.32	0.26	Included
Perillic aldehyde	150.22	11.82	11.66	0.15	Included
Proline	115.13	14.27	13.77	0.50	Excluded
Tetracosane	338.65	18.17	Not found	n/a	Not found

Table S3: Various standards used to create chemical mixtures that were spiked on gauze containing human sebum, were made by dissolving them in appropriate solvents. The table shows solvents in which each of these standards were individually created, before forming mixtures at various concentrations used for validation of smell by the Super Smeller.

Standard	Solvent
3,4-dihydroxy mandelic acid	Water
Cyclohexasiloxane, dodecamethyl	Methanol
Cyclohexylcyclohexane	Methanol
Diglycerol	Water
Dodecane	Ethanol
Eicosane	Acetone
Gallic acid ethyl ester	Water
Glutamine	Water
Hexyl acetate	MeOH
Hippuric Acid	Water
Hydroxymyristic acid	Dichloromethane
Octacosane	Chloroform
Octanal	Methanol
Perillic aldehyde	Water
Proline	Water

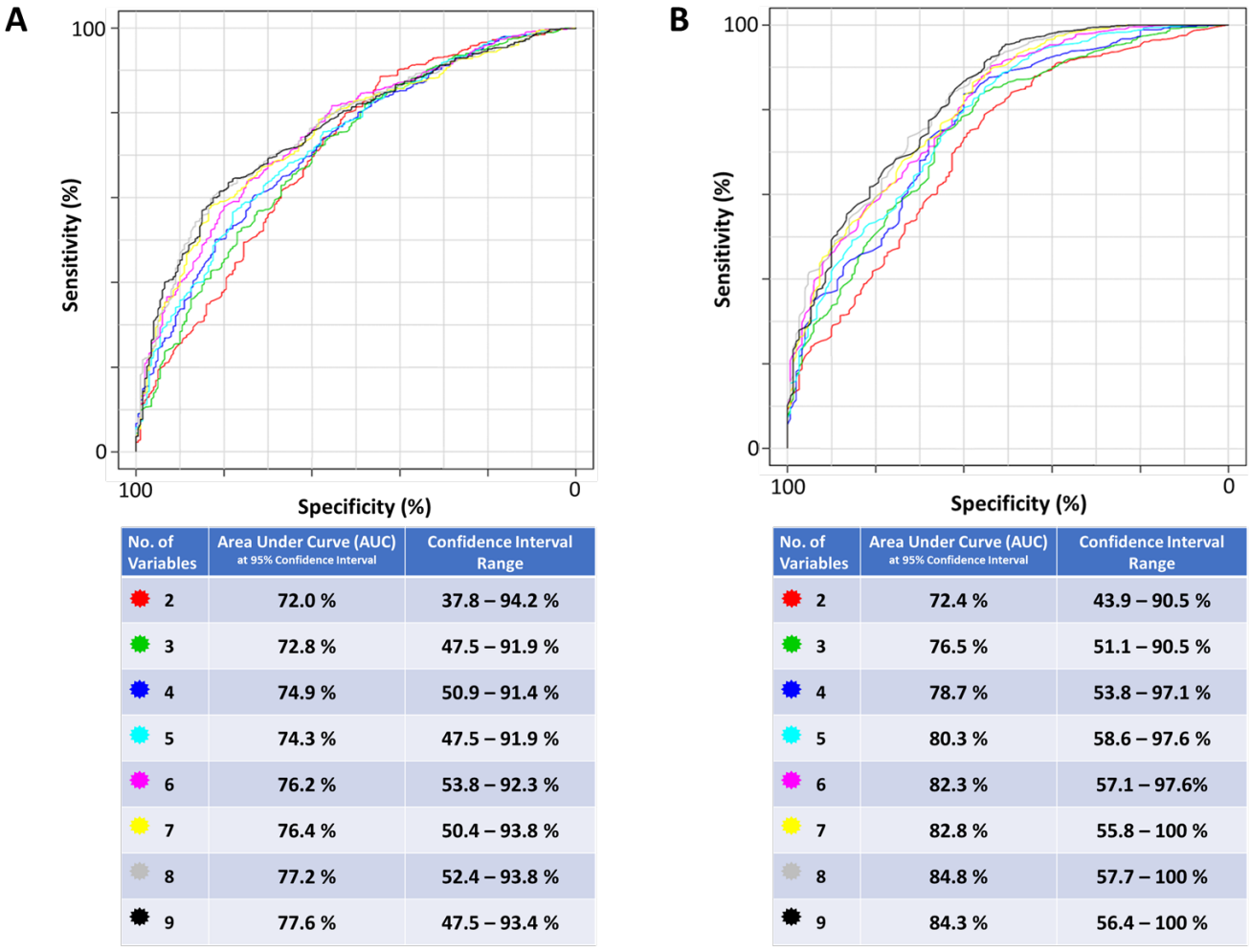
Table S4: Classification approaches attempted to distinguish between drug naïve PD participants and PD participants on medication did not show a very clear classification between the two groups in discovery cohort and validation cohort. *K*-nearest neighbours, random forest and support vector machines (SVM) classification algorithms were used. Using random sampling repeated 10 times, 60% data were used for training set and remaining data were used to test the model. Area under the curve (AUC) and averaged classification accuracy (ACA) for each model are shown.

	Discovery cohort		Validation cohort	
Method	**AUC**	**ACA**	**AUC**	**ACA**
kNN	78%	61%	57%	42%
SVM	65%	60%	60%	35%
Random Forest	66%	61%	54%	38%

Table S5: List of participants' anonymized ID along with time since they were diagnosed with Parkinson's as of the date of recruitment to this study. Median time since diagnosis for those in Drug Naïve group was 0 year whereas those in Medication group was 3 years.

ID	Year of diagnosis	Group	Time since diagnosis (years) when recruited
210717_005	2017	Drug Naive	0
210717_006	2017	Drug Naive	0
210717_016	2013	Drug Naive	4
210717_019	2017	Drug Naive	0
210717_022	2016	Drug Naive	1
210717_023	2015	Drug Naive	2
210717_025	2017	Drug Naive	0
210717_028	2013	Drug Naive	4
210717_030	2016	Drug Naive	1
210717_033	2017	Drug Naive	0
210717_038	2015	Drug Naive	2
181017_009	2017	Drug Naïve	0
181017_016	2017	Drug Naive	0
181017_019	2017	Drug Naive	0
181017_020	2005	Drug Naive	12
181017_021	2017	Drug Naive	0
181017_023	2017	Drug Naive	0
181017_024	2016	Drug Naive	1
191017_003	2016	Drug Naive	1
191017_004	2017	Drug Naive	0
191017_007	2015	Drug Naive	2
181017_004	2002	Medication	15
181017_005	2016	Medication	1
181017_006	2014	Medication	3
181017_010	2016	Medication	1
181017_012	2016	Medication	1
181017_013	2013	Medication	4
181017_018	2013	Medication	4
191017_006	2007	Medication	10
191017_008	2015	Medication	2
191017_009	2014	Medication	3

210717_009	2014	Medication	3
210717_012	2016	Medication	1
210717_018	2014	Medication	3
210717_029	2014	Medication	3
210717_031	2015	Medication	2
210717_032	2015	Medication	2
210717_034	2017	Medication	0
210717_035	2014	Medication	3
210717_037	2004	Medication	13



No. of Variables	Area Under Curve (AUC) at 95% Confidence Interval	Confidence Interval Range
2	72.0 %	37.8 – 94.2 %
3	72.8 %	47.5 – 91.9 %
4	74.9 %	50.9 – 91.4 %
5	74.3 %	47.5 – 91.9 %
6	76.2 %	53.8 – 92.3 %
7	76.4 %	50.4 – 93.8 %
8	77.2 %	52.4 – 93.8 %
9	77.6 %	47.5 – 93.4 %

No. of Variables	Area Under Curve (AUC) at 95% Confidence Interval	Confidence Interval Range
2	72.4 %	43.9 – 90.5 %
3	76.5 %	51.1 – 90.5 %
4	78.7 %	53.8 – 97.1 %
5	80.3 %	58.6 – 97.6 %
6	82.3 %	57.1 – 97.6%
7	82.8 %	55.8 – 100 %
8	84.8 %	57.7 – 100 %
9	84.3 %	56.4 – 100 %

Figure S1: ROC plots generated using all nine metabolites that were common between the two cohorts (A discovery and B validation) (but not necessarily differential using Student's t-test or expressed in the same direction between cohorts). Each model was built using PLS-DA to rank all variables and top two important variables were selected to start with. Then in each subsequent model additional variables by rank were added to generate ROC curve. Confidence intervals were

calculated by Monte Carlo Cross Validation (MCCV) using balanced sub-sampling with multiple repeats.

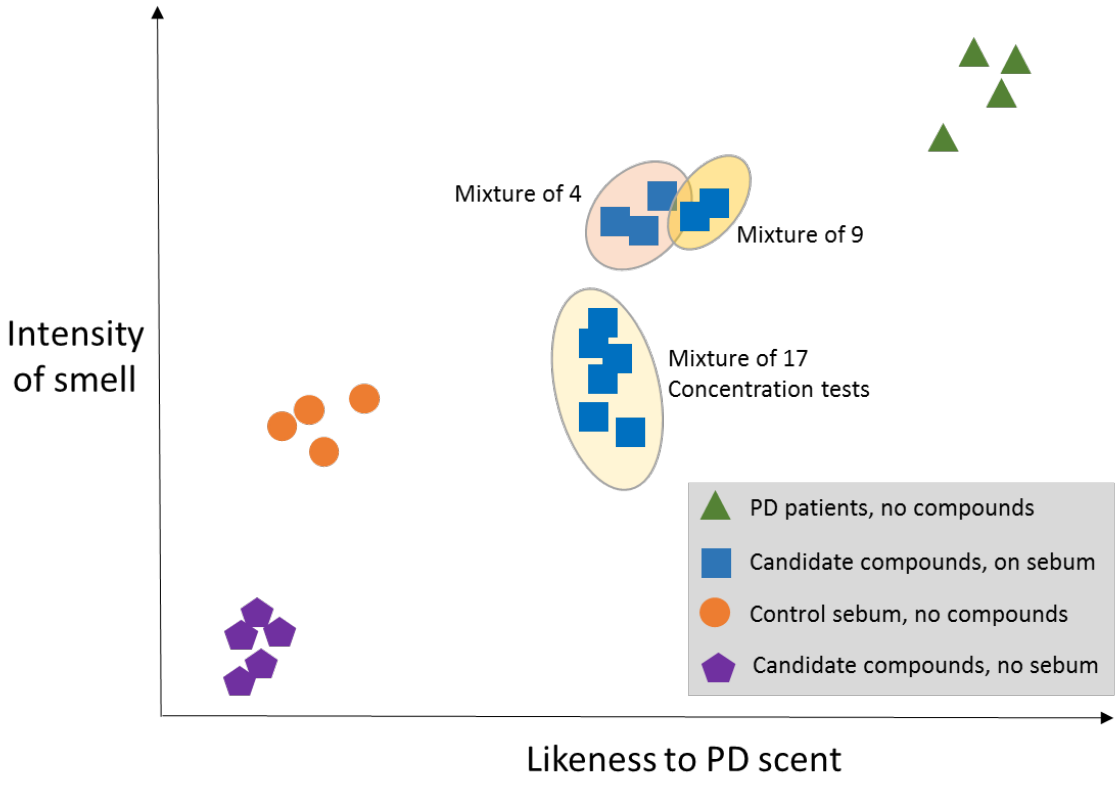


Figure S2: Schematic to show qualitatively the results of a series of blind randomised studies performed by the Super Smeller to classify and score samples based on their similarity to the 'PD smell' and their overall intensity of smell, these scores were defined by both oral and physical denomination by the Super Smeller. All samples were presented on gauze swabs; **purple** pentagons display a cluster of swabs of multiple combinations of candidate compounds spiked onto blank gauze (no sebum), **orange** circles show gauze swabs containing only human control sebum (no compounds), **blue** squares depict three series of compound combinations based on MS analysis spiked on to control sebum and **green** triangles represent clinical gauze samples swabbed from PD patients.