Metabolomics (2025) 21:17
https://doi.org/10.1007/s11306-024-02218-8

ORIGINAL ARTICLE

Optimized breath analysis: customized analytical methods and enhanced workflow for broader detection of VOCs

Wisenave Arulvasan[1] · Julia Greenwood[1] · Madeleine L. Ball[1] · Hsuan Chou[1] · Simon Coplowe[1] · Owen Birch[1] · Patrick Gordon[1] · Andreea Ratiu[1] · Elizabeth Lam[1] · Matteo Tardelli[1] · Monika Szkatulska[1] · Shane Swann[1] · Steven Levett[1] · Ella Mead[1] · Frederik-Jan van Schooten[2] · Agnieszka Smolinska[2] · Billy Boyle[1] · Max Allsworth[1]

Received: 1 November 2024 / Accepted: 31 December 2024

Abstract

Introduction Breath Volatile organic compounds (VOCs) are promising biomarkers for clinical purposes due to their unique properties. Translation of VOC biomarkers into the clinic depends on identification and validation: a challenge requiring collaboration, well-established protocols, and cross-comparison of data. Previously, we developed a breath collection and analysis method, resulting in 148 breath-borne VOCs identified.
Objectives To develop a complementary analytical method for the detection and identification of additional VOCs from breath. To develop and implement upgrades to the methodology for identifying features determined to be "on-breath" by comparing breath samples against paired background samples applying three metrics: standard deviation, paired t-test, and receiver-operating-characteristic (ROC) curve.
Methods A thermal desorption (TD)-gas chromatography (GC)-mass spectrometry (MS)-based analytical method utilizing a PEG phase GC column was developed for the detection of biologically relevant VOCs. The multi-step VOC identification methodology was upgraded through several developments: candidate VOC grouping schema, ion abundance correlation based spectral library creation approach, hybrid alkane-FAMES retention indexing, relative retention time matching, along with additional quality checks. In combination, these updates enable highly accurate identification of breath-borne VOCs, both on spectral and retention axes.
Results A total of 621 features were statistically determined as on-breath by at least one metric (standard deviation, paired t-test, or ROC). A total of 38 on-breath VOCs were able to be confidently identified from comparison to chemical standards.
Conclusion The total confirmed on-breath VOCs is now 186. We present an updated methodology for high-confidence VOC identification, and a new set of VOCs commonly found on-breath.

Keywords Volatile organic compounds (VOCs) · Breathomics · Microbiome · Volatile metabolites · Non-invasive biomarkers · VOC Atlas

1 Introduction

Volatile organic compounds (VOCs) contain at least one carbon atom and are emitted as gases from certain solids or liquids. In human metabolism, endogenously generated VOCs can cross most biological membranes away from their point of origin, enter the bloodstream, cross into the air in the lungs at the alveolar membrane, and be exhaled in the breath (Amann et al., 2014; Bax et al., 2019; Drabińska et al., 2021). Research has detected over 1400 VOCs in exhaled breath (Drabińska et al., 2021), many of which have great potential as non-invasive biomarkers for a variety of diseases such as cancer (Altomare et al., 2020; Bhandari et al., 2023; Hanna et al., 2019), liver disease (Dadamio et al., 2012; Ferrandino et al., 2020, 2023; Río et al., 2015), inflammatory bowel disease (IBD) (Bannaga et al., 2019; Henderson et al., 2022; Smolinska et al.,

✉ Wisenave Arulvasan
wisenave.arulvasan@owlstone.co.uk

[1] Owlstone Medical Ltd., Cambridge, UK

[2] Faculty of Health, Medicine and Life Sciences, Pharmacology and Toxicology, Maastricht University, Maastricht, Netherlands

Published online: 20 January 2025

2018), asthma (Azim et al., 2019; Dallinga et al., 2010; Djukanović et al., 2024), and more. A strongly emphasized point across the breath research literature is the lack of standardization or consistency of breath analysis techniques, chemical identification, background correction, and controls (Beauchamp et al., 2016; Chou et al., 2024a; Fiehn et al., 2007; Herbig & Beauchamp, 2014; Issitt et al., 2022; Jia et al., 2019; Pham et al., 2023; Schmidt et al., 2021; Spaněl et al., 2013; Summer et al., 2007). This has led to issues of reliability and lack of repeatability across studies, making the development of candidate biomarkers challenging (Chou et al., 2024b; Haworth et al., 2022). There is the added complication that chemical identification of VOCs detected in studies is not always accurate, and so the reporting of VOCs may appear to be inconsistent simply due to misidentified compounds. The highest confidence identification of a VOC requires a comparison to purified chemical standards analyzed using the same instrumentation and methods (Fiehn et al., 2007; Summer et al., 2007). At least one unique chemical standard is required for every VOC to be confidently identified, but as this is costly and time-consuming, this is not often undertaken.

The existence of a reference library of both chemically confirmed endogenous and exogenous VOC identities that genuinely originate from breath (and not from background contamination) will facilitate future VOC breath biomarker discovery and subsequent biomarker validation in clinical studies (Arulvasan et al., 2024). Additionally, this library of VOCs could be used to facilitate cross-study data comparisons for improved standardization across the field of breath research. To this end, a previously reported robust breath analysis platform led to the creation of an initial library of 148 "on-breath" VOCs that were quantifiably distinguishable from background contaminants (Arulvasan et al., 2024). These VOCs were identified using purified chemical standards in a heterogenous human population (Arulvasan et al., 2024). As over 1400 VOCs have been described in breath (Drabińska et al., 2021) (although many of these are likely to be inhaled contaminants from the environment), this is likely only a minority of the VOCs in exhaled breath that could serve as biomarkers. Therefore, expanding the methodology to reliably detect a wider range of breath VOCs is crucial, particularly those that can be linked to the gut microbiome and associated diseases that would benefit from breath-based diagnostics. This includes capturing VOCs such as p-cresol, 2-methylindole, and short-chain and branched-chain fatty acids (SCFAs, BCFAs) (Besten et al., 2013; Cagno et al., 2011; Gall et al., 2011; Ni et al., 2014; Rios-Covian et al., 2020; Swanson et al., 2002; Walton et al., 2013). Given the growing interest in the gut microbiome, broadening the detection of VOCs chemically classified as phenols, indoles and their derivatives, fatty acyls, and carboxylic acids and their derivatives would be especially beneficial to researchers.

Targeting biologically relevant VOCs for discovery using TD-GC-MS can be undertaken with a few key methodological expansions. To detect and identify specific VOCs of interest, it is crucial to tailor the GC column chemistry to the compounds' chemical properties. Some classes, such as fatty acids and phenylpropanoic acids, are polar due to their functional groups and therefore, suited for retention and separation by a polyethylene glycol (PEG) stationary phase GC column, particularly as it is acid deactivated, which will reduce the adsorption of acid compounds. In contrast, other target classes, such as naphthalenes are non-polar. Further adjustments to the method could lead to additional on-breath VOCs being retained, separated, and detected even if they are weakly polar or non-polar. This includes altering the operating temperature of the column. For example, the maximum operating temperature of TG-WaxMS A is 250 °C, in contrast to 320 °C for the TraceGOLDTM-624SilMS utilized in the previous study (Arulvasan et al., 2024). The lower operating temperature can reduce the decomposition of thermally labile VOCs and improve retention of lighter VOCs and thus, allow them to be detected. Lower GC temperatures can also reduce the elution speed of low-boiling point VOCs, meaning greater separation and thus, less co-elution of VOCs. The different phase chemistry of the column will also lead to different combinations of VOCs eluting, which can reduce or remove the incidences of co-elution.

In this study, a complementary analytical discovery method was developed, and the chemical identification (ID) workflow, which encompasses the process from sample analysis to robust identification of VOCs, was upgraded. To achieve this, a TG-WaxMS A column was utilized, and the analytical method parameters (temperatures and flows) were optimized accordingly for the resolved detection of target VOCs. The list of target VOCs was compiled and curated to include compounds commonly reported in literature and external databases as associated to biological processes, with particular emphasis on digestive health. The original on-breath VOC identification workflow was upgraded through multiple developments including an advanced in-house spectral library building process, schema for grouping of candidate VOC standards, and further quality control (QC) checks for increased ID confidence. Retention matching was also optimized further, through the development of a hybrid fatty acid methyl esters (FAMES)-alkane hybrid retention index (RI) ladder for accurate calculation of RI values for candidate VOCs and on-breath features for comparison. Relative retention time (RRT) tolerances were established to enable the identification of on-breath features eluting outside the range of the RI ladder. Where the candidate VOC is analyzed alongside breath samples in the same analytical sequence, additional tolerance thresholds for RT matching were also implemented, for increased identification confidence.

Analysis of breath and equipment background samples using the developed complementary analytical discovery method and upgraded ID workflow has resulted in the identification of 38 further on-breath VOCs, bringing the total number of identified on-breath VOCs to 186. This includes the aforementioned compounds, many of which hold clinical interest in the gut microbiome and its relevant diseases. This work is forming the foundation of building the Breath Biopsy VOC Atlas®, a catalog of identified and quantified volatile organic compounds (VOCs) originating from exhaled breath. The atlas is being built as a database, search interface, and analytics tool to describe high-confidence VOCs alongside their clinical, chemical, and biological context.

2 Methods

2.1 Breath sampling

The breath and background samples were collected using the Owlstone Medical Novel Insights (OMNI) method, which uses Owlstone Medical's ReCIVA® Breath Sampler, and CASPER Portable Air Supply® for breath collection as described previously (Arulvasan et al., 2024). For each breath sample taken, the ReCIVA sampler collects breath into four sorbent tubes, and two tubes undergo the analytical procedure. The paired background sample was collected immediately before each breath sample was taken. Because this study is a technical advancement of a previously published analytical method, the two remaining technical replicates from that study were utilized (Arulvasan et al., 2024). A total of 90 breath samples and 90 paired system background samples were analyzed in this study, from 90 adult subjects. It must be noted that the storage time is longer (median collection to analysis timespan of 480 days) than for the breath samples analyzed as part of the previous publication (Arulvasan et al., 2024). The samples were all collected within a 4 week time frame and stored in a dedicated temperature-controlled fridge at 5 °C. The average pre-purge storage time of samples was 7 days. The demographic distributions of the study participants are detailed in Supplementary Table 1.

2.2 Complementary analytical method development

A panel of target VOCs was initially defined (Supplementary Tables 7 and 8); this comprised partly of biologically relevant compounds belonging to chemical classes and subclasses that were not identified on-breath in the previous publication (Arulvasan et al., 2024) (examples include naphthalenes, fatty acids, phenylpropanoic acids, hydroxy acids and derivatives, cinnamic acids, dicarboxylic acids and derivatives). The panel also included bio-relevant compounds that were not identified on-method previously but belong to a chemical class that has been identified on-breath in the previous publication (Arulvasan et al., 2024). Different permutations of flow and temperature settings for the TD and GC were investigated for use with the TG-WaxMS A phase GC column,the configuration that demonstrated optimal analytical performance for separation and detection of target VOCs, was selected (Supplementary Table 2, Supplementary Fig. 1). Lower operating temperatures and different phase chemistry of the column will lead to a different combination of VOCs' ions entering the C-trap for injection into the orbitrap mass analyzer. The process is controlled by the AGC (automatic gain control) which monitors the incoming ion current by active feedback and limits the number of ions reaching the C-trap accordingly to prevent ion overloading. High concentrations VOCs, such as acetone, can saturate the C-trap capacity, leading to suppression of other lower concentration VOCs eluting from the GC at the same time, preventing their detection. By altering the order and timings of breath VOCs eluting from the column through changes in temperature and column chemistry, on-breath VOCs which were previously suppressed by large neighboring VOCs could be detected.

3 Breath analysis

Breath and background samples were analyzed through TD (Markes)—GC-MS (Q exactive Orbitrap (Thermo Fisher Scientific) high-resolution accurate mass spectrometry) platform using the developed complementary TG-WaxMS A analytical method (Supplementary Table 2). To enable the calculation of RI, FAMEs (C5–23) (Supplementary Table 5) and straight-chain alkanes (C5–C25) ladders were analyzed alongside (Castello, 1999). The alkane ladders were prepared using different solvents (C5–C16 (50 ng per alkane) solubilized in methanol, C17–C25 (20 ng per alkane) solubilized in cyclohexane) and spiked onto two separate tubes. To prevent MS detector overloading and ion source contamination from the larger and sustained cyclohexane solvent peak, a modified analytical method (deploying a filament delay) was set up to analyze the C17–C25 tube.

A five-point QC calibration curve of a subset (89 VOCs) of previously identified on-breath IDs (Arulvasan et al., 2024) that are detectable on the TG-Wax-MS A method were analyzed in each sequence for monitoring method and instrument performance; the compounds analyzed are listed in Supplementary Table 6.

Additionally, a panel of pre-defined on-method target candidate VOC standards previously not identified on-breath were analyzed in each sequence. These compounds spanned a range of chemical classes, including fatty acid esters, fatty alcohols, dicarboxylic acids and derivatives, carbonyl compounds, and monoterpenoids. The majority of these VOCs were analyzed as a four-point (0.1–50 ng) calibration panel (Supplementary Table 7) and the remaining at a single level (50 ng) Supplementary Table 8), due to the sequence capacity limit. Analyzing these VOCs alongside samples increased identification confidence, where a match is observed between the target VOC and an on-breath feature. To account for analytical variation, all samples were liquid injected with a mix of eight deuterated internal standard compounds dissolved in methanol (2.7 ng per I.S compound, except o-cresol-d8 and decane-d22: 12 and 9.25 ng, respectively) to result in comparable peak area signal as the other I.S compounds. Each paired breath and background samples were analyzed in the same sequence to reduce the technical variability, and a total of 18 sequences were run on the platform.

3.1 Post-processing: untargeted feature extraction, normalization & on-breath feature determination

The resulting breath and background samples' chromatograms were batch processed (signal deconvolution, feature group clustering, and library matching to NIST23), utilizing the OMNI untargeted feature extraction method in Compound Discoverer (ver. 3.3, Thermo Scientific™), detailed in Supplementary Table 3. After feature extraction, all data was normalized using the hybrid correlation-retention time method, detailed in the previous publication (Arulvasan et al., 2024). On-breath features were then statistically determined by the three metrics (standard deviation, paired t-test, and ROC-AUC) as previously described. NIST hits for features that were on-breath by at least one of the three metrics were curated and prioritized, considering match score (SI and RSI), hit ranking, and the number of metrics the feature was on-breath by. The selected candidate NIST hits were then further filtered for biological relevance by searching against literature and well-established databases, such as the human metabolome database (HMDB), small molecule pathway database (SMPDB), and Rhea, prior to sourcing the NIST hits' standards at a minimum of 95% purity. To ensure robust assessment, a systematic approach was employed including pathway involvement, source validation, and relevance to human physiology and disease, supported by curated databases and literature searches (Novoa-del-Toro & Witting, 2024).

3.2 Candidate VOC preparation and identification

Candidate VOCs originated from two sources: (1) a panel of 145 pre-defined target VOCs, analyzed alongside samples (Supplementary Tables 7 and 8) compromising primarily of biologically relevant on-method compounds of interest, and (2) prioritized biologically relevant NIST hits sourced for on-breath features generated during untargeted feature extraction of the samples analyzed on the TG-WAX-MS A analytical method, as detailed in the prior sections. To generate data efficiently and prevent misidentification, candidate VOCs were prepared into mixes prior to spiking onto TD tubes for analysis. Candidate VOCs were grouped together based on several criteria, including NIST polar RI difference (at least 50 units), molecular weight, chemical class, and chemical moiety, to reduce the likelihood of co-elution. To prevent intra-mix reactivity, which can generate unintended compound(s) leading to safety concerns and misidentification, chemical incompatibility guidance on the compound safety data sheets (SDS), and the guidelines stated in (Appendix, 2024) for grouping chemicals, were strictly adhered to. Acids and aldehyde candidate compounds were also not grouped together, due to the risk of ester formation. To ensure the peak corresponding to the spiked target is correctly identified, a forward and reverse similarity index (SI and RSI) score of at least 700 must be achieved to its corresponding NIST entry. Additionally, the peak signal intensity must be at least three times higher than the equivalent peak in the methanol blank analyzed within the same sequence. The specific isomer of a candidate VOC was analyzed where possible.

3.3 Spectral library building

Each candidate VOC was prepared and analyzed as a randomized ten-point calibration, at a wide range of concentrations (0.001–100 ng) to capture the linear dynamic range. After the exclusion of ions demonstrating poor peak shape by manual review of the deconvolved spectra, the data (m/z of ion and its measured abundance at each concentration level) for the candidate VOC peak was ingested by the in-house ion correlation tool, which calculates the correlation of each ion between each concentration level. The generated output contained the ions which exhibit a Spearman rank coefficient of 1 for abundance across at least three concentration levels (Fig. 1), as the final list of ions to include for building the in-house spectra for the candidate VOC. The concentration level containing the largest number of correlating ions was used to build the final in-house spectra.

To validate the quality advancement of our developed in-house library spectra-building approach, in-house spectra were created following the two methods (ion correlation approach detailed in the methods) and the original NIST

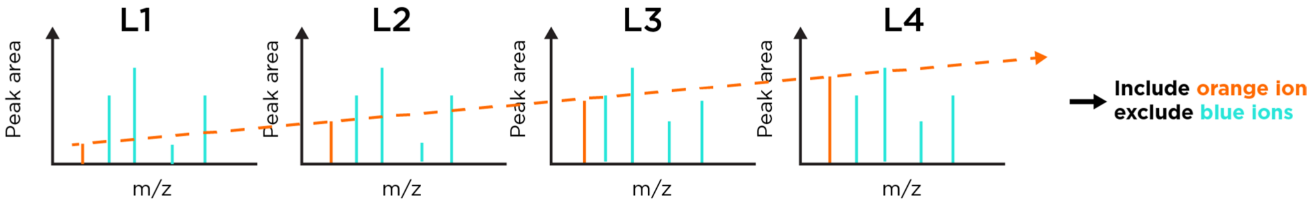


Fig. 1 Schematic illustrating the logic of the ion abundance correlation approach. Example spectra for 4 concentration levels of a compound is shown, with ions correlating (i.e. increased peak area response with increased concentration) across 4 levels (orange) and thus included in the in-house spectra for the compound, vs. Ions which do not exhibit correlation (blue)

spectra-driven approach described in the previous publication (Arulvasan et al., 2024) for a subset of on-breath ID VOC standards and compared.

3.4 ID confirmation: chromatographic retention matching metrics

3.4.1 Retention index ladders

Given our analytical method utilizes a polar GC column, polar compounds (such as fatty acid methyl ester (FAMEs), are expected to exhibit better compatibility with the column phase to produce sharp and symmetrical peaks, compared to alkanes and so, resulting in lower RT and thus, RI variability. Accurate retention index values minimize both false positive and negative VOC identifications. To achieve this, two chemical ladders- fatty acid methyl ester (FAMEs) and straight-chain alkanes- were analyzed within each analytical sequence for retention indices (RI) calculation. The FAMEs (C5–C23) (50 ng per compound) were solubilized in methanol and spiked onto one tube. The alkane ladder was extended since the previous publication (Arulvasan et al., 2024) from C16 to C25, to enable RI calculation of a greater proportion of on-breath features.

3.5 ID confirmation: spectral similarity matching

Along with retention similarity, spectral similarity between the candidate VOC in-house library spectra and the five breath samples with the highest abundance of the on-breath feature suspected to match, is calculated for assessing the match between a candidate VOC and on-breath feature. Match scores (SI and RSI) of 800 or higher for at least three of the five breath samples are required for a successful match.

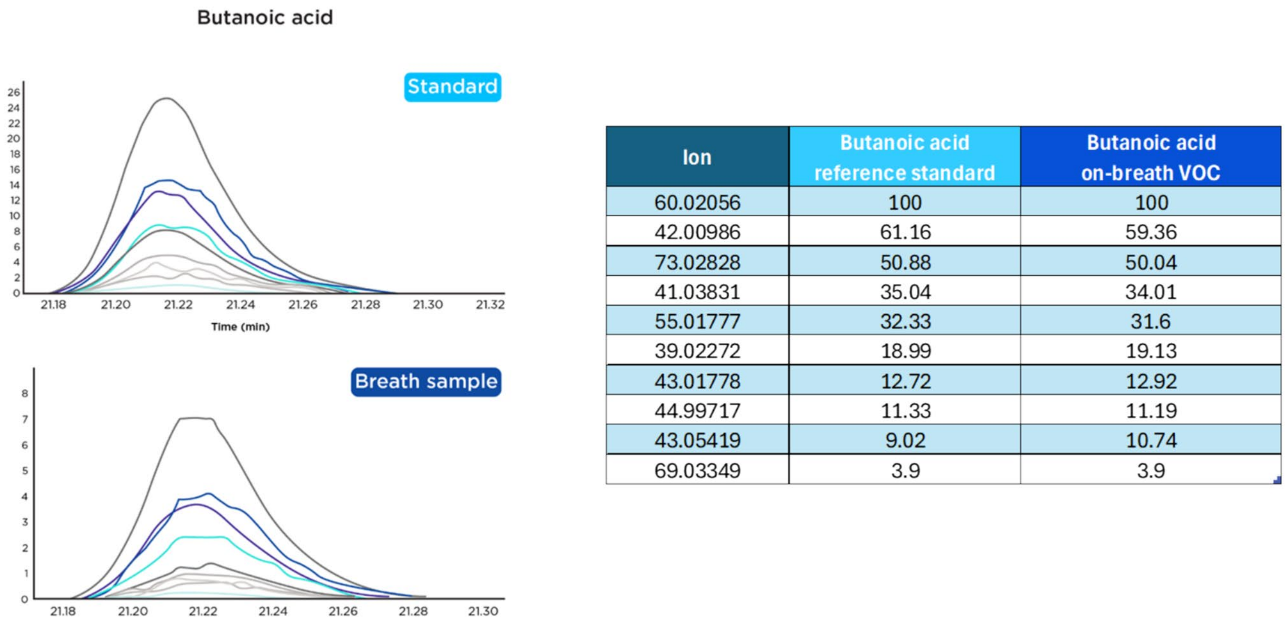
3.6 Additional quality checks

3.6.1 On-breath status validity check

During untargeted feature extraction, the Compound Discoverer algorithm selects a reference ion for each feature to minimize data sparsity. Thus, the ion most frequently observed across all deconvolved spectra, including both breath and background samples, is chosen by the software to extract the feature signal (peak area) for each sample. As these signal values are used to determine the on-breath status of a feature, it is crucial that the reference ion originates from the matching candidate VOC rather than a background ion or co-eluting feature, for the on-breath status to be valid for the matching VOC. To ensure this, the selected reference ion is compared against the ions present in the deconvolved in-house spectra of the matching candidate VOC. If the reference ion matches, the validation check is passed. If there is no match, the reference ion is manually altered to an ion derived from the candidate VOC, which subsequently alters the feature signal in each sample. These updated signal values are then used to recalculate the on-breath metrics. If the feature remains on-breath, the validation check is successful; if not, the feature-candidate VOC pair is excluded from on-breath identification.

3.6.2 Feature quality check

After an on-breath feature match with a candidate VOC is established by meeting the thresholds for a successful retention and spectral match, the quality of the peak was assessed, prior to final confirmation. To assess the quality of the candidate VOC and matching on-breath feature peak, the extracted ion chromatograms (EICs) for the ten most abundant ions in the in-house candidate VOC spectra were overlaid and compared with the corresponding EICs from the three breath samples containing the highest abundance for the matching on-breath feature, using FreeStyle™ version 1.8. On-breath feature identities were confirmed through their overlaid ion profiles. All confirmed identities displayed a Gaussian shape with good symmetry, which ensures no assignment to integrated noise or coeluting features. The visual comparison also allows us to determine how well the ion ratios match between breath and standard. Figure 2 illustrates the comparison of the 10 largest ions' ratios plotted for the standard (top) and its matching on-breath feature (bottom) for butanoic acid and tabulated alongside, demonstrating the



Ion	Butanoic acid reference standard	Butanoic acid on-breath VOC
60.02056	100	100
42.00986	61.16	59.36
73.02828	50.88	50.04
41.03831	35.04	34.01
55.01777	32.33	31.6
39.02272	18.99	19.13
43.01778	12.72	12.92
44.99717	11.33	11.19
43.05419	9.02	10.74
69.03349	3.9	3.9

Fig. 2 Left: Visual comparison of the extracted ion chromatograms (EICs) for the top 10 ions of the butanoic acid standard (top) and the matching on-breath feature (bottom). Matching ions are represented in the same color between the two EICs. Right: Tabulated 10 highest abundance ions' ratios calculated with respect to the base peak ion (m/z 60.02056) in the butanoic acid reference standard and the on-breath feature identified as butanoic acid in the highest abundance breath sample

presence of the same ions at very similar ratios between the two and thus, a successful match.

3.7 VOC identification workflow overview

The VOC identification workflow, displayed in Fig. 3 is a multi-step process from sample analysis to successful identification of on-breath features.

4 Results

4.1 Instrument QC

The five-point instrumental QC compounds' calibration curves analyzed alongside all sample sequences were processed utilizing targeted data processing in Chromeleon 7.3.2 MUb. Linear regression was applied and the resulting coefficient of determination (R^2) values for each QC compound calibration curve in each sequence was plotted as a series of boxplots (Supplementary Fig. 2). 99% of all R^2 values were above 95%, demonstrating high instrument & analytical method stability, and low random error.

4.2 Spectral library building

A novel, robust methodology, based on ion abundance correlation across concentration levels of the candidate VOC standard, has been developed for in-house spectral library construction, which was then matched against on-breath features. This approach offers several advantages, compared to the previously used reference library-centric approach, including the reduction of both type I and II errors during matching of an on-breath feature and the candidate VOC, through the removal of contaminant ions, while retaining truly relevant ions, such as those originating from the VOC and its interaction with the analytical flow path.

An interesting case study example is on-breath ID tetrachloroethylene, a vinyl halide: here, the ion with *m/z* 110.93998 has been included in the library spectra built using the ion correlation approach (as its abundance correlates with spiked concentration), however, excluded in the spectra built using original approach (given it does not have an associated theoretical fragment annotation and is absent in its NIST library spectra). This ion is also present in the deconvolved breath samples' on-breath feature confirmed to be tetrachloroethylene. While the exact mechanism through which the *m/z* 110.93998 ion is produced during Orbitrap Q exactive TD-GC-MS analysis of tetrachloroethylene is not yet established, evidence suggests this to be the product of a series of water gas phase reactions that occur in the C-trap, by the reaction with residual water in the system (Baumeister et al., 2019). A series of controlled experiments doping isotopically labeled water into the C-trap of an Orbitrap Q-exactive induced formation of additional ions during the analysis of halogen cyanides, not observed when conducting

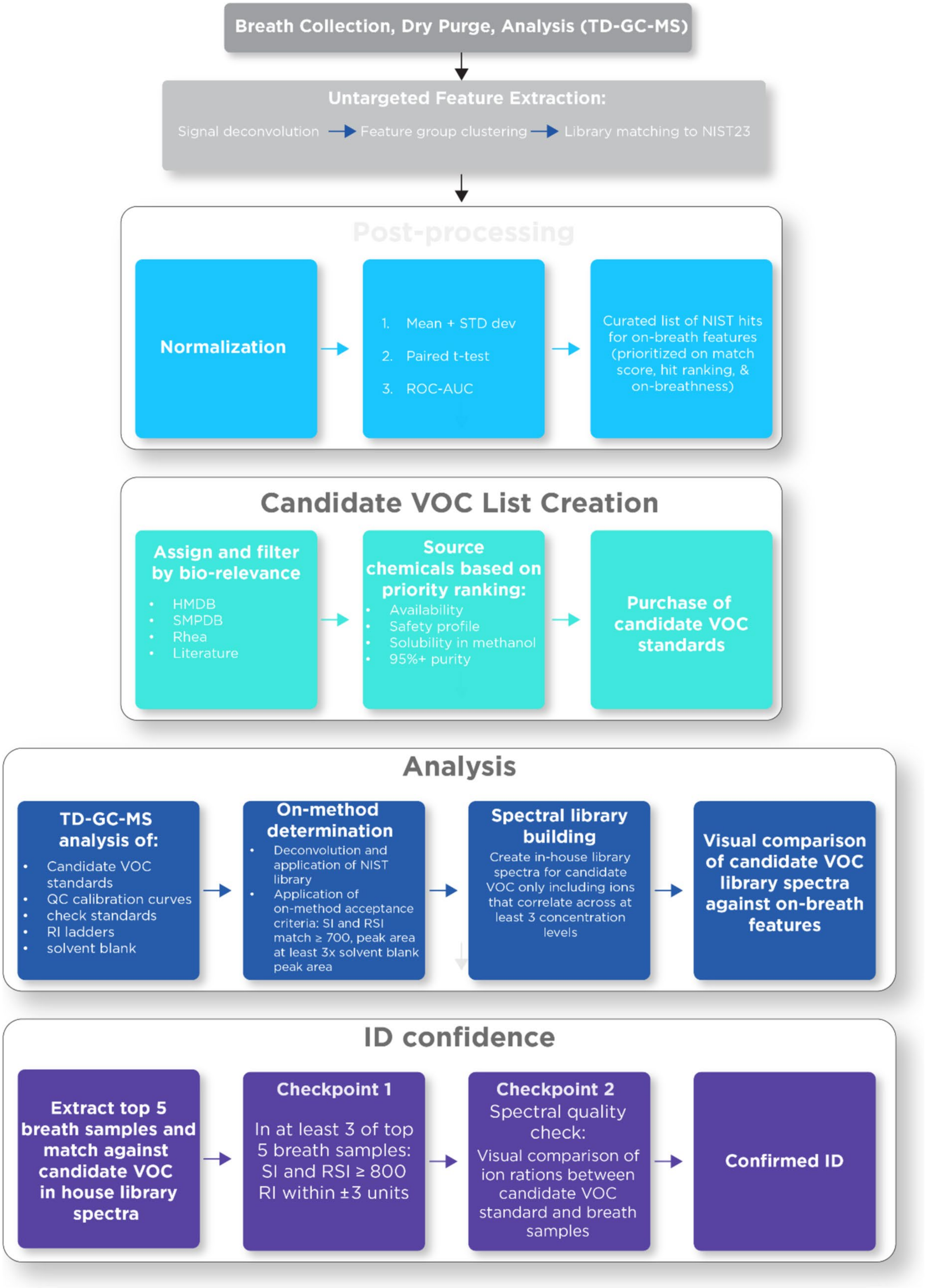

Breath Collection, Dry Purge, Analysis (TD-GC-MS)
Untargeted Feature Extraction:
Signal deconvolution → Feature group clustering → Library matching to NIST23
Post-processing
Normalization
1. Mean + STD dev
2. Paired t-test
3. ROC-AUC
Curated list of NIST hits for on-breath features (prioritized on match score, hit ranking, & on-breathness)
Candidate VOC List Creation
Assign and filter by bio-relevance
HMDB
SMPDB
Rhea
Literature
Source chemicals based on priority ranking:
Availability
Safety profile
Solubility in methanol
95%+ purity
Purchase of candidate VOC standards
Analysis
TD-GC-MS analysis of:
Candidate VOC standards
QC calibration curves
check standards
RI ladders
solvent blank
On-method determination
Deconvolution and application of NIST library
Application of on-method acceptance criteria: SI and RSI match ≥ 700, peak area at least 3x solvent blank peak area
Spectral library building
Create in-house library spectra for candidate VOC only including ions that correlate across at least 3 concentration levels
Visual comparison of candidate VOC library spectra against on-breath features
ID confidence
Extract top 5 breath samples and match against candidate VOC in house library spectra
Checkpoint 1
In at least 3 of top 5 breath samples: SI and RSI ≥ 800 RI within ±3 units
Checkpoint 2
Spectral quality check:
Visual comparison of ion rations between candidate VOC standard and breath samples
Confirmed ID

Fig. 3 Flow diagram summarizing the key steps of the end-to-end process of VOC identification from breath collection to confirmed on-breath VOC identity

the analysis on a single quadrupole instrument, which is the type of MS used to generate NIST spectra. Given the accurate mass capability of the Orbitrap Q exactive, the precise formula for the ion was calculated, which suggested it to be a product of hydrogen atomic transfer reaction, demonstrating the occurrence of gas phase reactions of analytes in the C-trap, in the presence of moisture (Gall et al., 2011).

Thus, in the case of the 110.93998 *m/z* ion fragment produced during the analysis of tetrachloroethylene (C_2Cl_4), it is hypothesized that this is formed through the reaction of the analyte with water to form C_2Cl_2OH (Fig. 4). The *m/z* value of C_2Cl_2OH (110.940446) matches the *m/z* value of the ion observed in the spectra of tetrachloroethylene in both breath and the standard (*m/z* 110.93998) after the removal of an electron. This case study demonstrates the benefit of utilizing ion abundance correlation to dictate which ions should be included in the in-house spectra for library matching, as this enables the inclusion of ions specifically produced by the analysis of a target VOC on our instrument. Without the inclusion of such ions, the resulting match score between the in-house library spectra and breath samples would be decreased and thus, the ion correlation approach to spectral building reduces the likelihood of false negative matches.

The final stages of ID confirmation rely upon meeting two key acceptance criteria: spectral and retention similarity between the candidate VOC standard and the on-breath feature. Therefore, ensuring the methodology for calculating and compensating for retention time drift and setting appropriate pass/fail thresholds, are crucial for ensuring reliable VOC identification.

4.3 Hybrid alkanes-FAMEs retention indexing approach development & validation

To enable higher accuracy of RI matching and increase the proportion of on-breath features that can be identified, the hybrid alkanes-FAMEs RI ladder was developed. This approach harnesses the advantage of the FAMEs ladder, which is theorized to produce more stable and consistent peaks with the TG WAX MS A polar column phase compared to alkanes, and the alkane ladder, whose retention timespan is wider than FAMEs, facilitating the calculation of RI for a greater number of on-breath features. The alkane ladder was used to calculate RI for features with a retention time (RT) below the earliest eluting FAMEs' (Methyl butyrate) and above the earliest eluting alkane (Pentane). The FAMEs ladder was used where the feature's RT is between the first and last eluting FAMEs (Methyl butyrate and Methyl behenate, respectively).

To experimentally validate this approach on our platform prior to implementation and define an RI tolerance for a successful on-breath feature–candidate VOC match, RI values were calculated for each QC and I.S compound, herein referred to as QC compounds (listed in Supplementary Tables 4 and 6) in each of the 18 sample analysis sequences by applying the non-isothermal Kovats retention index formula (NIST, 2024) using the three RI ladders (alkane only, FAMEs only, and the hybrid alkane-FAMEs ladder). The QC compounds spanned a range of chemical classes including fatty acid esters, indoles, alkanes and carbonyl compounds to ensure good generalizability of the resulting RI limit tolerances for matching chemically diverse candidate VOCs.

For the central 80% of the intra-batch QC RI data, the median absolute deviation (MAD) was calculated, which was then converted to an RI tolerance limit. This was done through multiplication by a scaling factor (for comparability to standard deviation) followed by multiplication by 3 to include 99.73% of the QC data, resulting in RI tolerance limits of 2.70 and 2.91 for the FAMEs only and the hybrid alkane-FAMEs ladder, respectively, compared to the larger tolerance of 3.99 for the alkane-only ladder. This experimentally demonstrated reduced RI variability when utilizing a ladder primarily based on FAMEs, as theorized. The density plot (Fig. 5) illustrates the intra-batch RI range for

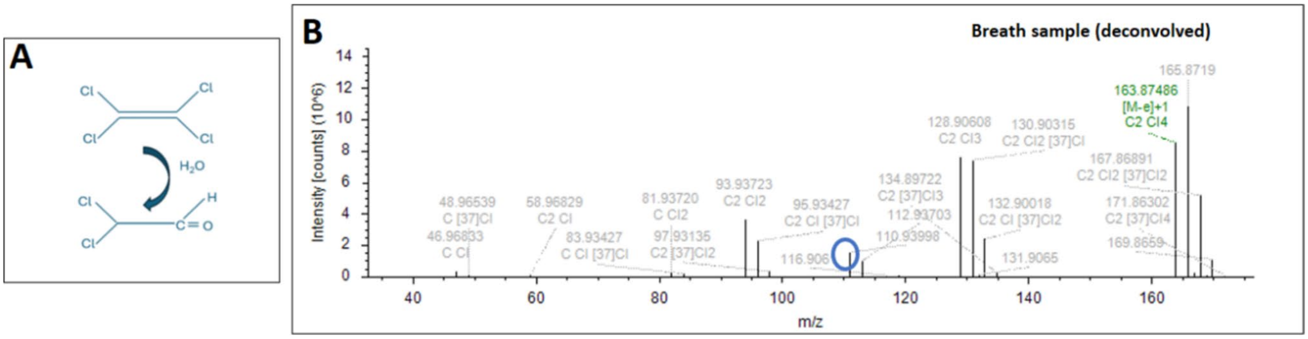


Fig. 4 **A** Chemical reaction schematic showing the hypothesized reaction of tetrachloroethylene with water to produce the C_2Cl_2OH fragment. **B** Deconvolved breath sample peak matching tetrachloroethylene illustrating the presence of the m/z 110.93998 ion, with m/z value on the x-axis and peak intensity (counts) on the y-axis

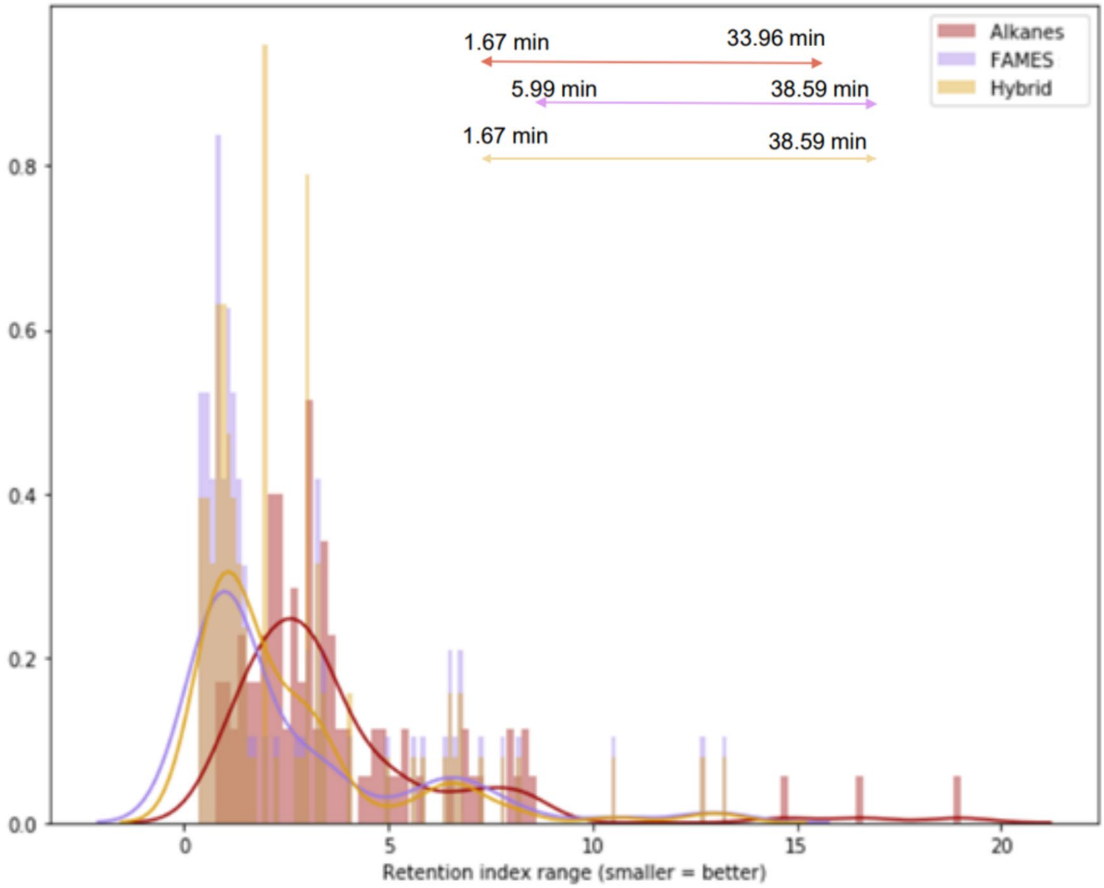


Fig. 5 Density plot summarizing the proportion of the QC compounds (y-axis) vs. calculated retention index range (x-axis) across the 18 sample sequences, using retention index ladders of alkane only (red), fatty acid methyl ester (FAMEs) only (lilac) and the hybrid alkane-FAMEs (yellow)

each QC compound calculated using the three RI ladders; on average, the RI range is smaller for the FAMEs and hybrid ladders, than the alkanes, indicating reduced RI variability with ladders containing FAMEs for our analytical method.

Although the RI variability was slightly lower for the FAMEs only ladder than the hybrid ladder, the hybrid ladder encapsulates a larger proportion of the chromatogram, enabling calculation of RI values for 8% more on-breath features compared to using FAMEs only, whilst maintaining similarly low RI variability as the FAMEs ladder. This means that the RI approach can be used for identifying a larger proportion of on-breath features and the risk of false positive matches are reduced due to greater stability of RI values for the compound between breath samples and the standards analysis, often several sequences later. The 2.91 units RI threshold was rounded up to be ± 3 RI units as the acceptance limit for RI matching, on-breath features to candidate VOCs, using the hybrid alkane-FAMEs ladder approach.

4.4 Relative retention time (RRT) and retention time (RT)

To enable the identification of on-breath features that fall outside of the hybrid retention index ladder boundaries, relative retention time (RRT) was utilized. RRT was calculated by normalizing the RT of the candidate VOC/on-breath feature to the closest eluting spiked internal standard compound's RT within the same tube. For a successful retention match between an on-breath feature and candidate VOC, the acceptance limit was ± 0.002 units. This was calculated using a similar approach to the RI tolerance calculation described in the Methods. A subset of biologically relevant candidate VOCs were pre-defined and analyzed alongside the samples (Supplementary Tables 7 and 8). For such candidate VOCs, an additional stringent RT-based matching criteria for retention similarity was applied for higher feature identification confidence. An RT match within ± 4 s, between the on-breath feature in at least three of the five breath samples containing the highest signal, and the candidate VOC analyzed in the same sequence, was deemed acceptable. The full VOC identification workflow utilized in this study, from sample analysis to successful identification of on-breath VOCs is shown in Fig. 3.

4.5 Identified VOCs

A total of 1775 features were identified, and from these, 621 features were statistically determined as on-breath by at least one metric, with 223 (36%) being classified as on-breath by all three metrics. Metric 1 applied a flexible frequency threshold at 3 standard deviations above background, with a 50% threshold used in this study, resulting in 254 molecular features. This threshold, chosen to capture features prevalent in most samples, can be adjusted for different analytical requirements. A more stringent threshold would lead to fewer features classified as on-breath. The analytical methods and chemical identification (ID) workflow described above resulted in the confident identification of 38 on-breath VOCs using chemical standards not previously identified in the previous publication (Arulvasan et al., 2024) (Table 1). Two further features were assigned identities of triethylene glycol monoethyl ether and levoglucosenone, however, these VOCs were only considered on-breath by type 1 in 15 and 10 breath samples, respectively, and thus would not meet the 50% on-breath threshold specified for this metric.

Twenty-five on-breath IDs are on-breath by all three metrics (standard deviation (metric 1), paired t-test (metric 2), and ROC-AUC (metric 3), as described in Sect. 2.4 of the previous publication (Arulvasan et al., 2024) Six and three IDs are on-breath solely by metric 2 and 3, respectively. Figure 6 illustrates the total number of on-breath features detected using three distinct calculation metrics, highlighting significant overlap in the subsets classified by each.

5 Discussion

On-breath VOCs represent a group of compounds above background that have been confidently chemically identified, and reliably detected in exhaled breath of a heterogenous population of human volunteers (Arulvasan et al., 2024). Previously, 148 on-breath VOCs were identified, and in this study, the expanded analytical workflow enabled the detection and identification of 38 more. A complementary analytical method utilizing a different GC column was developed and implemented. Along with this, several optimizations to the ID workflow including establishing a hybrid alkanes-FAMEs retention indexing ladder, enhanced in-house spectral library building methodology and additional quality control checks, have contributed to expanding the range of VOC chemistries that can be added to the growing Breath Biopsy VOC Atlas resource, ensuring a more comprehensive coverage of VOC profiles.

To identify specific biologically relevant VOCs for further biomarker studies and applications, it is crucial to tailor the GC column chemistry to the chemical properties of the compounds of interest. One limitation of the TraceGOLD™-624SilMS GC column in our previous study was that some on-breath VOCs from under-represented chemical classes were not effectively retained and separated. The selection of the TG-WaxMS A phase GC column in this study overcomes some of these limitations and provides several advantages: (i) the PEG phase is highly polar, making it ideal for the retention and resolved separation of polar VOCs, such as acids, alcohols, esters, and aldehyde, (ii) the TG-WaxMS A column is specifically treated to analyze acidic polar compounds, such as free fatty acids, carboxylic acids, which are biologically relevant, without the need for derivatization. The TG-WaxMS A phase GC column enhances compound sensitivity and minimizes interference, facilitating the detection and identification of ultra-low concentration on-breath VOCs, and improving match scores between breath samples and standards by reducing background ions. While the TG-WAX-MS-A column has demonstrated suitability for the detection of several bio-relevant on-breath VOCs, it is important to emphasize that column chemistry choice should be driven by the physiochemical properties of the target compound(s). Therefore, detection of other bio-relevant VOCs may require the utilization of alternative GC column chemistries. In this study, a hybrid alkanes-FAMEs retention indexing approach was developed and validated to harness both the breadth and specificity advantages of alkane and FAMEs ladders, respectively. The FAMEs ladder is representative of matching key chemical classes, including fatty acids, and compatible, in terms of polarity, with the TG Wax MS A column phase, while the alkane ladder is suitable for calculating RI for candidate VOCs that elute before the earliest FAMEs' RT. Compared to the previous publication, which only encompassed the alkane ladder to C16, the extension to C25 in this study increases the proportion of the sample and candidate VOC chromatograms encapsulated by the retention index ladder. Here the latest NIST version (NIST23) was used for untargeted feature extraction, expanding the range of candidate VOCs for consideration during feature extraction given the inclusion of over 87,000 additional compound spectra. This enhances the comparison against on-breath features and the availability of RI data. The robust methodology developed to build an in-house spectral library, based on ion abundance correlation, offers several key advantages. First, it provides additional assurance in selecting the correct peak corresponding to the injected target VOC standard, as the peak area will positively correlate with concentration levels. Additionally, it reduces false positives in on-breath feature identification by

Table 1 Table showing all 38 on-breath VOCs that were able to be confidently chemically identified in this study, along with their InChI key identifier and chemical classification

On-breath VOC ID	InChI key	Class	Subclass	% of breath samples on-breath in by metric 1	On-breath status		
					Metric 1 (mean + std dev)	Metric 2 (paired t-test)	Metric 3 (ROC-AUC)
2-Methylimidazole	LXBGSDVWAMZHDD-UHFFFAOYSA-N	Azoles	Imidazoles	46	FALSE	TRUE	FALSE
2-Isopropenyltoluene	OGMSGZZPTQNTIK-UHFFFAOYSA-N	Benzene and substituted derivatives	Phenylpropenes	82	TRUE	TRUE	TRUE
Salicylhydrazide	XSXYESVZDBAKKT-UHFFFAOYSA-N	Benzene and substituted derivatives	Benzoic acids and derivatives	22	FALSE	FALSE	TRUE
1-benzofuran-2-carbonitrile	ZQGAXHXHVKVERC-UHFFFAOYSA-N	Benzofurans		10	FALSE	FALSE	TRUE
Acetamide	DLFVBJFMPXGRIB-UHFFFAOYSA-N	Carboximidic acids and derivatives	Carboximidic acids	79	TRUE	TRUE	TRUE
Isobutyric acid	KQNPFQTWMSNSAP-UHFFFAOYSA-N	Carboxylic acids and derivatives	Carboxylic acids	81	TRUE	TRUE	TRUE
Chlorodibromoacetic acid	UCZDDMGNCJJAHK-UHFFFAOYSA-N	Carboxylic acids and derivatives	Alpha-halocarboxylic acids and derivatives	23	FALSE	TRUE	FALSE
N,N-Dimethylacetamide	FXHOOIRPVKKKFG-UHFFFAOYSA-N	Carboxylic acids and derivatives	Carboxylic acid derivatives	56	TRUE	TRUE	TRUE
2-Propenoic acid	NIXOWILDQLNWCW-UHFFFAOYSA-N	Carboxylic acids and derivatives	Acrylic acids and derivatives	96	TRUE	TRUE	TRUE
4-Nitrophenyl 2-(2-(((benzyloxy)carbonyl)amino)acetamido)acetate	VJPZTTLCVFONEM-UHFFFAOYSA-N	Carboxylic acids and derivatives	Amino acids, peptides, and analogues	53	TRUE	TRUE	TRUE
2-Acetoxycinnamic Acid	UXOWQQCLBQBRMQ-VOTSOKGWSA-N	Cinnamic acids and derivatives	Hydroxycinnamic acids and derivatives	50	TRUE	TRUE	TRUE
Butyric acid	FERIUCNNQQJTOY-UHFFFAOYSA-N	Fatty Acyls	Fatty acids and conjugates	54	TRUE	TRUE	TRUE
Isovaleric Acid	GWYFCOCPABKNJV-UHFFFAOYSA-N	Fatty Acyls	Fatty acids and conjugates	55	TRUE	TRUE	TRUE
4-Methylvaleric acid	FGKJLKRYENPLQH-UHFFFAOYSA-N	Fatty Acyls	Fatty acids and conjugates	73	TRUE	TRUE	TRUE
2-Methylbutyric acid	WLAMNBDJUVNPJU-UHFFFAOYSA-N	Fatty Acyls	Fatty acids and conjugates	77	TRUE	TRUE	TRUE
Isovaleramide	SANOUVWGPVYVAV-UHFFFAOYSA-N	Fatty Acyls	Fatty amides	92	TRUE	TRUE	TRUE
Methyl 2-methylbutyrate	OCWLYWIFNDCWRZ-UHFFFAOYSA-N	Fatty Acyls	Fatty acid esters	6	FALSE	FALSE	TRUE

Table 1 (continued)

On-breath VOC ID	InChI key	Class	Subclass	% of breath samples on-breath in by metric 1	On-breath status		
					Metric 1 (mean + std dev)	Metric 2 (paired t-test)	Metric 3 (ROC-AUC)
3-Methylfuran	KJRRQXYWFQKJIP-UHFFFAOYSA-N	Heteroaromatic compounds		99	TRUE	TRUE	TRUE
Furan	YLQBMQCUIZJEEH-UHFFFAOYSA-N	Heteroaromatic compounds		62	TRUE	TRUE	TRUE
2-Methylthiophene	XQQBUAPQHNYYRS-UHFFFAOYSA-N	Heteroaromatic compounds		79	TRUE	TRUE	TRUE
2-Methylindole	BHNHHSOHWZKFOX-UHFFFAOYSA-N	Indoles and derivatives	Indoles	58	TRUE	TRUE	TRUE
Phthalimide	XKJCHHZQLQNZHY-UHFFFAOYSA-N	Isoindoles and derivatives	Isoindolines	8	FALSE	TRUE	FALSE
Methyl isothiocyanate	LGDSHSYDSCRFAB-UHFFFAOYSA-N	Isothiocyanates		95	TRUE	TRUE	TRUE
Triethylsilanol	WVMSIBFANXCZKT-UHFFFAOYSA-N	Organometalloid compounds	Organosilicon compounds	42	FALSE	TRUE	FALSE
Acrolein	HGINCPLSRVDWNT-UHFFFAOYSA-N	Organooxygen compounds	Carbonyl compounds	37	FALSE	TRUE	TRUE
2-Heptanone	CATSNJVOTSVZJV-UHFFFAOYSA-N	Organooxygen compounds	Carbonyl compounds	68	TRUE	TRUE	TRUE
2-Methyl-2-cyclopenten-1-one	ZSBWUNDRDHVNJL-UHFFFAOYSA-N	Organooxygen compounds	Carbonyl compounds	27	FALSE	TRUE	TRUE
Propionaldehyde	NBBJYMSMWIIQGU-UHFFFAOYSA-N	Organooxygen compounds	Carbonyl compounds	26	TRUE	TRUE	TRUE
p-Cresol	IWDCLRJOBJJRNH-UHFFFAOYSA-N	Phenols	Cresols	68	TRUE	TRUE	TRUE
Terpinolene	MOYAFQVGZZPNRA-UHFFFAOYSA-N	Prenol lipids	Monoterpenoids	83	TRUE	TRUE	TRUE
Carvone	ULDHMXUKGWMISQ-UHFFFAOYSA-N	Prenol lipids	Monoterpenoids	58	TRUE	TRUE	TRUE
Menthone isomer (Menthone or Isomenthone*)	NFLGAXVYCFJBMK-BDAKNGLRSA-N	Prenol lipids	Monoterpenoids	69	TRUE	TRUE	TRUE
Menthone isomer (Menthone or Isomenthone*)	NFLGAXVYCFJBMK-BDAKNGLRSA-N	Prenol lipids	Monoterpenoids	73	TRUE	TRUE	TRUE
Menthol	NOOLISFMXDJSKH-UHFFFAOYSA-N	Prenol lipids	Monoterpenoids	31	FALSE	TRUE	TRUE

Table 1 (continued)

On-breath VOC ID	InChI key	Class	Subclass	% of breath samples on-breath in by metric 1	On-breath status		
					Metric 1 (mean + std dev)	Metric 2 (paired t-test)	Metric 3 (ROC-AUC)
α-Terpinene	YHQGMYU-VUMAZJR-UHFFFAOYSA-N	Prenol lipids	Monoterpenoids	97	TRUE	TRUE	TRUE
Linalool	CDOSHBSSF-JOMGT-UHFF-FAOYSA-N	Prenol lipids	Monoterpenoids	6	FALSE	TRUE	FALSE
6,6-Dimethylbicyclo[3.1.1]hept-2-ene-2-carbaldehyde	KMRMUZ-KLFIEVAO-UHFFFAOYSA-N	Prenol lipids	Monoterpenoids	12	FALSE	TRUE	FALSE
Dimethyl Sulfone	HHVIBTZHL-RERCL-UHFF-FAOYSA-N	Pyrans	Pyranones and derivatives	96	TRUE	TRUE	TRUE

The criteria they passed (TRUE) and failed (FALSE) to be classified as on-breath by the three metrics is also shown. The compounds are sorted by alphabetical order of chemical class

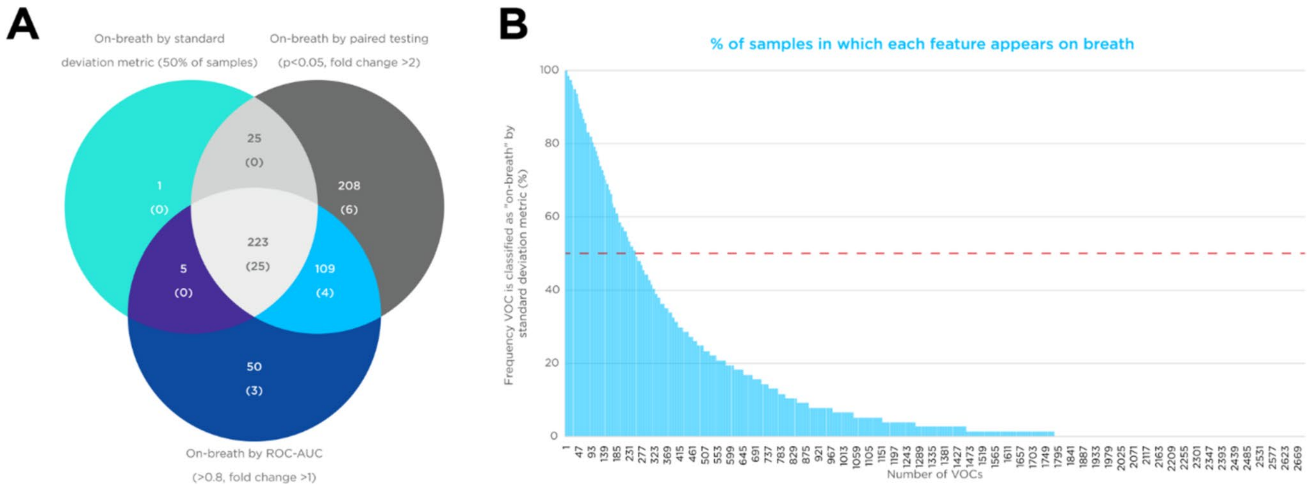


Fig. 6 **A** Venn diagram showing the numbers of VOCs classified as on-breath by each metric, along with the number of those VOCs that have been identified, in brackets. **B** Bar chart showing the frequency with which individual VOCs are classified as on-breath across all samples using metric 1. The dotted line indicates the 50% threshold, restricting the number of on-breath VOCs above this cut-off to 254 of the 1775 total

excluding ions originating from instrument background or introduced during pre-analytical and analytical processes. Finally, including ions generated by the VOCs' interactions with the TD-GC-MS analytical instrument flow path helps minimize false negatives in on-breath feature identification.

About three-quarters of the 38 additional on-breath VOCs identified in this study are highly relevant for clinical biomarker research, including some of the most well-known and influential volatile metabolites produced by the gut microbiome, which are presented in Fig. 7. SCFAs and BCFAs are both products of microbial fermentation. Previously, SCFAs acetic and propionic acid were identified as on-breath, but butyric acid was not (Arulvasan et al., 2024). The current workflow was able to classify butyric acid and BCFAs, such as isobutyric acid, isovaleric acid, 2-methylbutyric acid, and 4-methylvaleric acid, as on-breath VOCs.

Studies have shown that as humans age, there is a negative correlation between BCFAs and fiber intake, suggesting a microbiome decline (Rios-Covian et al., 2020). Interestingly, while SCFAs and BCFAs are both products of microbial fermentation, they seem to exhibit divergent trends (Tuck et al., 2020). Although the actual mechanism driving

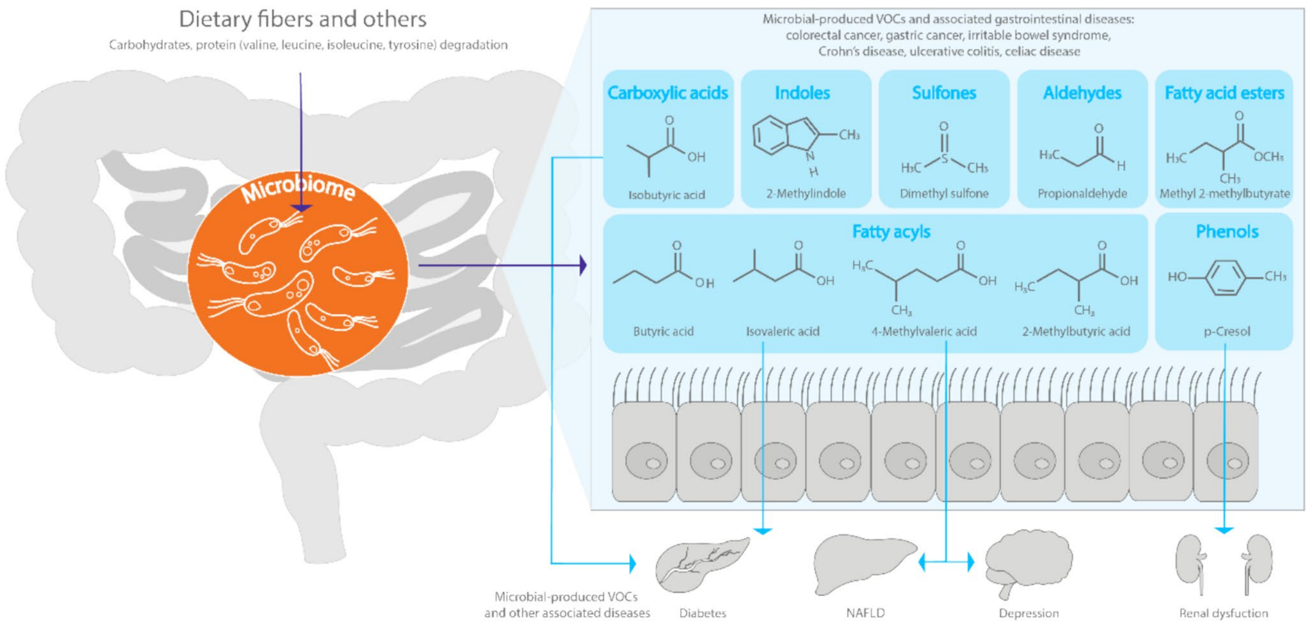


Fig. 7 Illustration of the origins of microbial-produced VOCs, their chemical (sub)classes, and associated diseases

these different fermentation patterns remains unknown, BCFAs—specifically those identified using our updated methods—have been linked to gut-microbiome-associated gastrointestinal diseases (Fig. 7) (Averina et al., 2020; Cagno et al., 2011; Gall et al., 2011; Gao et al., 2022; Garner et al., 2007; Goedert et al., 2014; Lin et al., 2021; Preter et al., 2015; Raman et al., 2013; Sinha et al., 2016; Walton et al., 2013; Weir et al., 2013).Together, these studies suggest that the gut microbiome plays a key role in regulating intestinal permeability, chronic systemic inflammation, and blood–brain barrier permeability.

Additional VOCs identified in this study that are thought to be produced by microbes include p-cresol, 2-methylindole, dimethyl sulfone, propionaldehyde, and methyl 2-methylbutyrate (Fig. 7). p-Cresol has been associated with colorectal cancer (Ni et al., 2014), IBS (Walton et al., 2013), and UC (Gall et al., 2011), and is an end-product of tyrosine breakdown by intestinal bacteria, primarily aerobes. 2-Methylindole has been shown to affect bowel function via fructo-oligosaccharides and *Lactobacillus acidophilus* (Swanson et al., 2002). Dimethyl sulfone (DMSO2), associated with UC and CD (Dawiskiba et al., 2014), is derived from dietary sources, intestinal bacterial metabolism, and human endogenous methanethiol metabolism. Propionaldehyde has been linked to gastric cancer progression and can be produced by hydrolyzing propane-1,2-diol, an alternative pathway of propionate synthesis presented in several major microbiota phylogenetic groups, particularly *Lachnospiraceae* (Krishnan et al., 2015). Methyl 2-methylbutyrate may also be microbiome-related, as it has been implicated in gastrointestinal diseases like UC and CD (Ahmed et al., 2016).

There are a few limitations in this study. Although additional on-breath features were captured and identified using these new methods, a relatively small proportion of the total on-breath features were identified. The ability to identify more features was constrained by the lack of available standards and the safety profile of the VOCs, along with some of the on-breath features indicated to be artifacts. Also, the on-breath VOCs previously assigned identities (Arulvasan et al., 2024) were not re-identified in this dataset to focus efforts on identifying new on-breath features. Despite obtaining candidate VOC standards with ≥ 95% purity and applying stringent peak deconvolution parameters during feature extraction, there is still a possibility that the target VOC peak contains ions originating from co-eluting impurities. Therefore, this library-building method may not effectively filter out such ions, as they are likely to scale with concentration. However, the risk of this is expected to be low, given the high purity of standards used, and any impurities would need to have a very similar RT to the target VOC to be considered for inclusion. Additionally, the method remains limited in its ability to detect certain biologically relevant chemical classes, such as dicarboxylic acids and certain aldehydes. Compound derivatization may offer a promising solution (Pietrogrande et al., 2010), though it will require extensive further development. Nonetheless, the newly developed method addresses some of the drawbacks of the previous publication, which relied on ions from the NIST reference spectra for mass spectral cleaning.

In terms of pre-analytical factors, long term storage of samples could impact the compounds held on sorbent in two ways: the amount of compound could decrease (due to degradation or egress from the tubes) or the amount of compound could increase (due to the compound being a degradation product or contaminant ingress into the tubes). The utilization of on-breath metrics within our study protects against false-positives i.e. misidentification of features representing contamination ingress, which would happen equally to breath and background samples. However, degradation does introduce a risk of false negatives (i.e. not identifying VOCs that had once been on-breath). Whilst this is a limitation of the study, it does not invalidate the reported findings.

If the amount of compound on-tube increases because the compound is a degradation product then this may happen proportionally to the amount of signal on the tubes, and so this case does lead to the potential for misclassification of a VOC as on-breath. The rate of VOC breakdown on sorbent is compound-dependent: it may happen almost instantly or over long periods of time. This is therefore always a risk with regards to breath VOC identification involving capture onto sorbent, but we do acknowledge that this risk is increased with the longer storage time. Longitudinal storage stability data, prior to identification of the VOC, would not have the ability to distinguish between whether the tracked "on-breath" feature(s) are unaltered (truly breath borne) compound(s) or degradation product(s), as a result of a reaction occurring during sampling or shortly afterwards. Therefore, storage stability data would not help us mitigate the risk of misclassification.

Confident identification of on-breath VOCs, as described in this paper, provides the foundation onto which future targeted research could be applied to characterize and subsequently optimize pre-analytical factors, such as storage, dry-purge and inter-tube precision, specifically for the compounds that are known to truly originate from breath. Future work will focus on quantifying these identified VOCs and further analytical modifications, such as capturing new on-breath VOCs using different sorbent tubes, to expand the total number of on-breath VOCs. Another relevant aspect will be validation protocols for on-breath VOCs as potential biomarkers. These protocols should focus on defining specificity and sensitivity of the on-breath biomarkers, which could enhance the clinical applicability of VOCs identified, especially for those with potential diagnostic value. Sensitivity testing should consider the matrix effects, of breath samples, especially given that some VOCs may be present at concentrations near the instrument's limit of detection. Once established, validation protocols testing biomarker reproducibility across different instruments, laboratories, and sample cohorts would be essential to demonstrate consistent detection and quantification.

6 Conclusion

Previously, we reported a methodology capable of reliably detecting 148 on-breath VOCs. The updated analytical method and chemical identification workflow presented here have increased confidence in identifying on-breath VOCs, leading to 38 additional biologically relevant VOCs being confidently identified in the same cohort. Notably, 98% of these compounds are present in the Human Metabolome Database (HMDB), highlighting the promise of leveraging the non-invasive nature of breath sampling and VOC biomarkers to inform underlying physiological processes for various clinical purposes. Translating breath biomarker tests into clinical practice depends on the confident identification and validation of breath VOCs, a challenge that requires close collaboration, well-established protocols, and the ability to compare and validate data. The expansion of chemically identified on-breath VOCs in the Atlas database would significantly enhance the development and validation of breath VOC biomarker for clinical use.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11306-024-02218-8.

Author contributions JG, FvS, and WA designed the study. WA, SC, AR, EL, EM, MS, SL, SS and MT performed the research. WA, SC, and OB analyzed the data. WA, HC, MB, and JG wrote the paper. JG, WA, MB, HC, PG, AS, BB, and MA critically reviewed and revised the paper.

Data availability Data is provided within the manuscript or supplementary information files.

Declarations

Competing interests BB is the CEO and Co-Founder of Owlstone Medical Ltd. WA, HC, JG, MB, OB, SC, PG, AR, EL, MT, MS, SL, EM, CCP, LND, SS, and MA are, or were employees of Owlstone Medical Ltd at the time this work was completed.

Ethical approval All procedures performed in this study involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants included in the study, in accordance with ICH GCP. The study supporting this work was approved by the Reading Independent Ethics Committee RIEC: 290620-1.

References

Ahmed, I., Greenwood, R., Costello, B., Ratcliffe, N., & Probert, C. S. (2016). Investigation of faecal volatile organic metabolites as novel diagnostic biomarkers in inflammatory bowel disease. *Alimentary Pharmacology & Therapeutics, 43*(5), 596–611.

Altomare, D. F., Picciariello, A., Rotelli, M. T., De Fazio, M., Aresta, A., Zambonin, C. G., Vincenti, L., Trerotoli, P., & De Vietro, N. (2020). Chemical signature of colorectal cancer: Case–control study for profiling the breath print. *BJS Open, 4*(6), 1189–1199.

Amann, A., de Lacy Costello, B., Miekisch, W., Schubert, J., Buszewski, B., Pleil, J., Ratcliffe, N., & Risby, T. (2014). The human volatilome: Volatile organic compounds (VOCs) in exhaled breath, skin emanations, urine, feces and saliva. *Journal of Breath Research, 8*(3), 034001.

Appendix K. (2024). Incompatible chemicals. Environment, Health and Safety. Retrieved October 21, 2024, from https://ehs.cornell.edu/research-safety/chemical-safety/laboratory-safety-manual/appendix-k-incompatible-chemicals

Arulvasan, W., Chou, H., Greenwood, J., Ball, M. L., Birch, O., Coplowe, S., Gordon, P., Ratiu, A., Lam, E., Hatch, A., & Szkatulska, M. (2024). High-quality identification of volatile organic compounds (VOCs) originating from breath. *Metabolomics, 20*(5), 102.

Averina, O. V., Zorkina, Y. A., Yunes, R. A., Kovtun, A. S., Ushakova, V. M., Morozova, A. Y., Kostyuk, G. P., Danilenko, V. N., & Chekhonin, V. P. (2020). Bacterial metabolites of human gut microbiota correlating with depression. *International Journal of Molecular Sciences, 21*(23), 9234.

Azim, A., Barber, C., Dennison, P., Riley, J., & Howarth, P. (2019). Exhaled volatile organic compounds in adult asthma: A systematic review. *European Respiratory Journal*. https://doi.org/10.1183/13993003.00056-2019

Bannaga, A. S., Farrugia, A., & Arasaradnam, R. P. (2019). Diagnosing Inflammatory bowel disease using noninvasive applications of volatile organic compounds: A systematic review. *Expert Review of Gastroenterology & Hepatology., 13*(11), 1113–1122.

Baumeister, T. U. H., Ueberschaar, N., & Pohnert, G. (2019). Gas-phase chemistry in the GC orbitrap mass spectrometer. *Journal of the American Society for Mass Spectrometry, 30*(4), 573–580.

Bax, C., Lotesoriere, B. J., Sironi, S., & Capelli, L. (2019). Review and comparison of cancer biomarker trends in urine as a basis for new diagnostic pathways. *Cancers, 11*(9), 1244.

Beauchamp, J., Pleil, J., Risby, T., & Dweik, R. (2016). Report from IABR breath summit 2016 in Zurich, Switzerland. *Journal of Breath Research, 10*(4), 049001.

Bhandari, M. P., Polaka, I., Vangravs, R., Mezmale, L., Veliks, V., Kirshners, A., Mochalski, P., Dias-Neto, E., & Leja, M. (2023). Volatile markers for cancer in exhaled breath—could they be the signature of the gut microbiota? *Molecules, 28*(8), 3488.

Castello, G. (1999). Retention index systems: Alternatives to the *n*-alkanes as calibration standards. *Journal of Chromatography a., 842*(1), 51–64.

Chou, H., Godbeer, L., & Ball, M. L. (2024a). Establishing breath as a biomarker platform—take home messages from the Breath Biopsy Conference 2023. *Journal of Breath Research*. https://doi.org/10.1088/1752-7163/ad3fdf

Chou, H., Godbeer, L., Allsworth, M., Boyle, B., & Ball, M. L. (2024b). Progress and challenges of developing volatile metabolites from exhaled breath as a biomarker platform. *Metabolomics, 20*(4), 72.

Dadamio, J., Van den Velde, S., Laleman, W., Van Hee, P., Coucke, W., Nevens, F., & Quirynen, M. (2012). Breath biomarkers of liver cirrhosis. *Journal of Chromatography, 15*(905), 17–22.

Dallinga, J. W., Robroeks, C. M. H. H. T., Van Berkel, J. J. B. N., Moonen, E. J. C., Godschalk, R. W. L., Jöbsis, Q., Dompeling, E., Wouters, E. F. M., & Van Schooten, F. J. (2010). Volatile organic compounds in exhaled breath as a diagnostic tool for asthma in children. *Clinical & Experimental Allergy, 40*(1), 68–76.

Dawiskiba, T., Deja, S., Mulak, A., Ząbek, A., Jawień, E., Pawełka, D., et al. (2014). Serum and urine metabolomic fingerprinting in diagnostics of inflammatory bowel diseases. *World Journal of Gastroenterology, 20*(1), 163–174.

De Preter, V., Machiels, K., Joossens, M., Arijs, I., Matthys, C., Vermeire, S., Rutgeerts, P., & Verbeke, K. (2015). Faecal metabolite profiling identifies medium-chain fatty acids as discriminating compounds in IBD. *Gut, 64*(3), 447–458.

Del Río, R. F., O'Hara, M. E., Holt, A., Pemberton, P., Shah, T., Whitehouse, T., & Mayhew, C. A. (2015). Volatile biomarkers in breath associated with liver cirrhosis—comparisons of pre-and post-liver transplant breath samples. *eBioMedicine, 2*(9), 1243–1250.

den Besten, G., van Eunen, K., Groen, A. K., Venema, K., Reijngoud, D. J., & Bakker, B. M. (2013). The role of short-chain fatty acids in the interplay between diet, gut microbiota, and host energy metabolism. *Journal of Lipid Research, 54*(9), 2325–2340.

Di Cagno, R., De Angelis, M., De Pasquale, I., Ndagijimana, M., Vernocchi, P., Ricciuti, P., Gagliardi, F., Laghi, L., Crecchio, C., Guerzoni, M. E., & Gobbetti, M. (2011). Duodenal and faecal microbiota of celiac children: Molecular, phenotype and metabolome characterization. *BMC Microbiology., 11*(1), 219.

Djukanović, R., Brinkman, P., Kolmert, J., Gomez, C., Schofield, J., Brandsma, J., Shapanis, A., Skipp, P. J., Postle, A., Wheelock, C., & Dahlen, S. E. (2024). Biomarker predictors of clinical efficacy of the anti-IgE biologic omalizumab in severe asthma in adults: Results of the SoMOSA study. *American Journal of Respiratory and Critical Care Medicine, 210*(3), 288–297.

Drabińska, N., Flynn, C., Ratcliffe, N., Belluomo, I., Myridakis, A., Gould, O., Fois, M., Smart, A., Devine, T., & Costello, B. D. L. (2021). A literature survey of all volatiles from healthy human breath and bodily fluids: The human volatilome. *Journal of Breath Research, 15*(3), 034001.

Ferrandino, G., Orf, I., Smith, R., Calcagno, M., Thind, A. K., Debiram-Beecham, I., Williams, M., Gandelman, O., de Saedeleer, A., Kibble, G., & Lydon, A. M. (2020). Breath biopsy assessment of liver disease using an exogenous volatile organic compound—toward improved detection of liver impairment. *Clinical and Translational Gastroenterology, 11*(9), e00239.

Ferrandino, G., De Palo, G., Murgia, A., Birch, O., Tawfike, A., Smith, R., Debiram-Beecham, I., Gandelman, O., Kibble, G., Lydon, A. M., & Groves, A. (2023). Breath Biopsy® to Identify Exhaled Volatile Organic Compounds Biomarkers for Liver Cirrhosis Detection. *Journal of Clinical and Translational Hepatology, 11*(3), 638.

Fiehn, O., Robertson, D., Griffin, J., van der Werf, M., Nikolau, B., Morrison, N., Sumner, L. W., Goodacre, R., Hardy, N. W., Taylor, C., & Fostel, J. (2007). The metabolomics standards initiative (MSI). *Metabolomics, 3*, 175–178.

Gao, Y., Chen, H., Li, J., Ren, S., Yang, Z., Zhou, Y., & Xuan, R. (2022). Alterations of gut microbiota-derived metabolites in gestational diabetes mellitus and clinical significance. *Journal of Clinical Laboratory Analysis, 36*(4), e24333.

Garner, C. E., Smith, S., de Lacy Costello, B., White, P., Spencer, R., Probert, C. S., & Ratcliffem, N. M. (2007). Volatile organic compounds from feces and their potential for diagnosis of gastrointestinal disease. *The FASEB Journal, 21*(8), 1675–1688.

Goedert, J. J., Sampson, J. N., Moore, S. C., Xiao, Q., Xiong, X., Hayes, R. B., Ahn, J., Shi, J., & Sinha, R. (2014). Fecal metabolomics: Assay performance and association with colorectal cancer. *Carcinogenesis, 35*(9), 2089–2096.

Hanna, G. B., Boshier, P. R., Markar, S. R., & Romano, A. (2019). Accuracy and methodologic challenges of volatile organic compound-based exhaled breath tests for cancer diagnosis: A systematic review and meta-analysis. *JAMA Oncology, 5*(1), e182815.

Haworth, J. J., Pitcher, C. K., Ferrandino, G., Hobson, A. R., Pappan, K. L., & Lawson, J. L. D. (2022). Breathing new life into clinical testing and diagnostics: Perspectives on volatile biomarkers from breath. *Critical Reviews in Clinical Laboratory Sciences., 59*(5), 353–372.

Henderson, B., Meurs, J., Lamers, C. R., Batista, G. L., Materić, D., Bertinetto, C. G., Bongers, C. C., Holzinger, R., Harren, F. J., Jansen, J. J., & Hopman, M. T. (2022). Non-invasive monitoring of inflammation in inflammatory bowel disease patients during prolonged exercise via exhaled breath volatile organic compounds. *Metabolites, 12*(3), 224.

Herbig, J., & Beauchamp, J. (2014). Towards standardization in the analysis of breath gas volatiles. *Journal of Breath Research, 8*(3), 037101.

Issitt, T., Wiggins, L., Veysey, M., Sweeney, S. T., Brackenbury, W. J., & Redeker, K. (2022). Volatile compounds in human breath: Critical review and meta-analysis. *Journal of Breath Research, 16*(2), 024001.

Jia, Z., Patra, A., Kutty, V. K., & Venkatesan, T. (2019). Critical review of volatile organic compound analysis in breath and in vitro cell culture for detection of lung cancer. *Metabolites, 9*(3), 52.

Krishnan, S., Alden, N., & Lee, K. (2015). Pathways and functions of gut microbiota metabolism impacting host physiology. *Current Opinion in Biotechnology., 2*(36), 137.

Le Gall, G., Noor, S. O., Ridgway, K., Scovell, L., Jamieson, C., Johnson, I. T., Colquhoun, I. J., Kemsley, E. K., & Narbad, A. (2011). Metabolomics of fecal extracts detects altered metabolic activity of gut microbiota in ulcerative colitis and irritable bowel syndrome. *Journal of Proteome Research, 10*(9), 4208–4218.

Lin, H., Guo, Q., Wen, Z., Tan, S., Chen, J., Lin, L., Chen, P., He, J., Wen, J., & Chen, Y. (2021). The multiple effects of fecal microbiota transplantation on diarrhea-predominant irritable bowel syndrome (IBS-D) patients with anxiety and depression behaviors. *Microbial Cell Factories, 20*, 233.

Ni, Y., Xie, G., & Jia, W. (2014). Metabonomics of human colorectal cancer: New approaches for early diagnosis and biomarker discovery. *Journal of Proteome Research, 13*(9), 3857–3870.

NIST. (2024). Gas chromatographic retention data. Retrieved October 21, 2024, from https://webbook.nist.gov/chemistry/gc-ri/

Novoa-del-Toro, E. M., & Witting, M. (2024). Navigating common pitfalls in metabolite identification and metabolomics bioinformatics. *Metabolomics, 20*(5), 103.

Pham, Y. L., Holz, O., & Beauchamp, J. (2023). Emissions and uptake of volatiles by sampling components in breath analysis. *Journal of Breath Research, 17*(3), 037102.

Pietrogrande, M. C., Bacco, D., & Mercuriali, M. (2010). GC–MS analysis of low-molecular-weight dicarboxylic acids in atmospheric aerosol: Comparison between silylation and esterification derivatization procedures. *Analytical and Bioanalytical Chemistry, 396*(2), 877–885.

Raman, M., Ahmed, I., Gillevet, P. M., Probert, C. S., Ratcliffe, N. M., Smith, S., Greenwood, R., Sikaroodi, M., Lam, V., Crotty, P., & Bailey, J. (2013). Fecal microbiome and volatile organic compound metabolome in obese humans with nonalcoholic fatty liver disease. *Clinical Gastroenterology and Hepatology., 11*(7), 868-875.e3.

Rios-Covian, D., González, S., Nogacka, A. M., Arboleya, S., Salazar, N., Gueimonde, M., & de Los Reyes-Gavilán, C. G. (2020). An overview on fecal branched short-chain fatty acids along human life and as related with body mass index: Associated dietary and anthropometric factors. *Frontiers in Microbiology, 11*, 973.

Schmidt, A. J., Salman, D., Pleil, J., Thomas, C. L. P., & Davis, C. E. (2021). IABR Symposium 2021 meeting report: Breath standardization, sampling, and testing in a time of COVID-19. *Journal of Breath Research, 16*(1), 010201.

Sinha, R., Ahn, J., Sampson, J. N., Shi, J., Yu, G., Xiong, X., Hayes, R. B., & Goedert, J. J. (2016). Fecal microbiota, fecal metabolome, and colorectal cancer interrelations. *PLoS ONE, 11*(3), e0152126.

Smolinska, A., Tedjo, D. I., Blanchet, L., Bodelier, A., Pierik, M. J., Masclee, A. A., Dallinga, J., Savelkoul, P. H., Jonkers, D. M., Penders, J., & van Schooten, F. J. (2018). Volatile metabolites in breath strongly correlate with gut microbiome in CD patients. *Analytica Chimica Acta, 1025*, 1–11.

Spaněl, P., Dryahina, K., & Smith, D. (2013). A quantitative study of the influence of inhaled compounds on their concentrations in exhaled breath. *Journal of Breath Research, 7*(1), 017106.

Summer, L. W., Amberg, A., Barrett, D., Beale, M. H., Beger, R., Daykin, C. A., Fan, T. W. M., Fiehn, O., Goodacre, R., Griffin, J. L., & Hankemeier, T. (2007). Proposed minimum reporting standards for chemical analysis. *Metabolomics, 3*(3), 211–221.

Swanson, K. S., Grieshop, C. M., Flickinger, E. A., Bauer, L. L., Wolf, B. W., Chow, J., Garleb, K. A., Williams, J. A., & Fahey, G. C., Jr. (2002). Fructooligosaccharides and *Lactobacillus acidophilus* modify bowel function and protein catabolites excreted by healthy humans. *The Journal of Nutrition, 132*(10), 3042–3050.

Tuck, C. J., De Palma, G., Takami, K., Brant, B., Caminero, A., Reed, D. E., Muir, J. G., Gibson, P. R., Winterborn, A., Verdu, E. F., & Bercik, P. (2020). Nutritional profile of rodent diets impacts experimental reproducibility in microbiome preclinical research. *Scientific Reports, 10*(1), 17784.

Walton, C., Fowler, D. P., Turner, C., Jia, W., Whitehead, R. N., Griffiths, L., Dawson, C., Waring, R. H., Ramsden, D. B., Cole, J. A., & Cauchi, M. (2013). Analysis of volatile organic compounds of bacterial origin in chronic gastrointestinal diseases. *Inflammatory Bowel Diseases, 19*(10), 2069–2078.

Weir, T. L., Manter, D. K., Sheflin, A. M., Barnett, B. A., Heuberger, A. L., & Ryan, E. P. (2013). Stool microbiome and metabolome differences between colorectal cancer patients and healthy adults. *PLoS ONE, 8*(8), e70803.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.