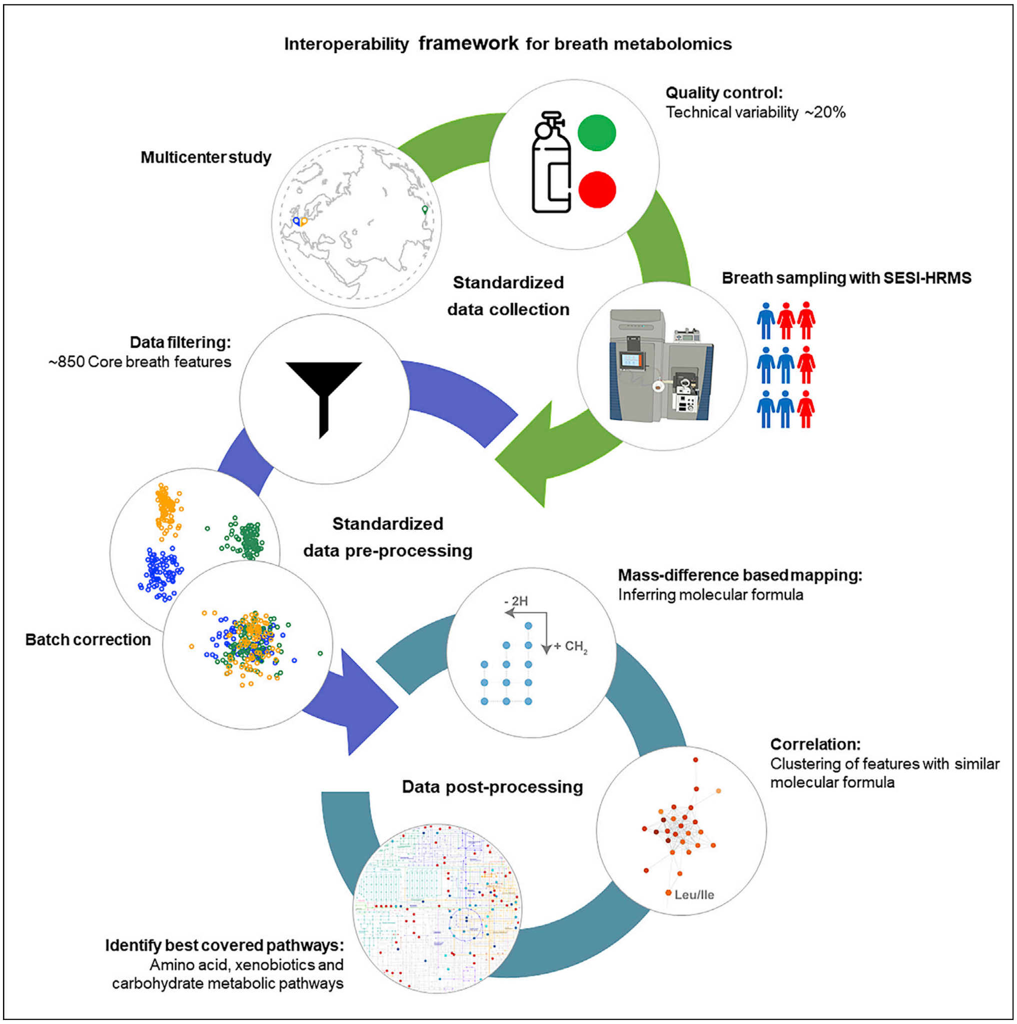
iScience

CellPress
OPEN ACCESS

Article

An interoperability framework for multicentric breath metabolomic studies


Interoperability framework for breath metabolomics
Multicenter study
Quality control:
Technical variability ~20%
Breath sampling with SESI-HRMS
Standardized
data collection
Data filtering:
~850 Core breath features
Standardized
data pre-processing
Batch correction
Mass-difference based mapping:
Inferring molecular formula
- 2H
+ CH₂
Correlation:
Clustering of features with similar molecular formula
Leu/Ile
Data post-processing
Identify best covered pathways:
Amino acid, xenobiotics and carbohydrate metabolic pathways

Amanda Gisler, Kapil Dev Singh, Jiafa Zeng, ..., Malcolm Kohler, Xue Li, Pablo Sinues

pablo.sinues@unibas.ch

Highlights

Technical variability of a SESI-HRMS instrument is ~20%

Human exhaled metabolome largely consists of ~850 core breath features

Core breath features are predominantly mapped to amino acid metabolism pathways

Interoperability framework paves the way for future large-scale multicenter trials

Gisler et al., iScience *25*, 105557
December 22, 2022 © 2022 The Authors.
https://doi.org/10.1016/j.isci.2022.105557

iScience

CellPress
OPEN ACCESS

Article

An interoperability framework for multicentric breath metabolomic studies

Amanda Gisler,[1,5] Kapil Dev Singh,[1,2,5] Jiafa Zeng,[1,2,3] Martin Osswald,[4] Mo Awchi,[1,2] Fabienne Decrue,[1] Felix Schmidt,[4] Noriane A. Sievi,[4] Xing Chen,[3] Jakob Usemann,[1] Urs Frey,[1] Malcolm Kohler,[4] Xue Li,[3] and Pablo Sinues[1,2,6,*]

SUMMARY

Exhaled breath contains valuable information at the molecular level and offers promising potential for precision medicine. However, few breath tests transition to routine clinical practice, partly because of the missing validation in multicenter trials. Therefore, we developed and applied an interoperability framework for standardized multicenter data acquisition and processing for breath analysis with secondary electrospray ionization-high resolution mass spectrometry. We aimed to determine the technical variability and metabolic coverage. Comparison of multicenter data revealed a technical variability of ~20% and a core signature of the human exhaled metabolome consisting of ~850 features, corresponding mainly to amino acid, xenobiotic, and carbohydrate metabolic pathways. In addition, we found high inter-subject variability for certain metabolic classes (e.g., amino acids and fatty acids), whereas other regions such as the TCA cycle were relatively stable across subjects. The interoperability framework and overview of metabolic coverage presented here will pave the way for future large-scale multicenter trials.

INTRODUCTION

Deep phenotyping of blood, urine, and tissue specimens with multiple -omics platforms has become increasingly important for clinical decision making, patient stratification, and therapeutic drug monitoring in modern medicine.[1] In striking contrast, exhaled breath remains largely excluded from the pool of exploited biofluids towards the hallmark of precision medicine.[2] Notable exceptions include the fractional exhaled nitric oxide (FeNO) test, used to diagnose and guide asthma management,[3,4] and the ^{13}C urea breath test, used to diagnose *Helicobacter pylori* infection.[5] While the biochemical richness of human breath was revealed over 50 years ago by Pauling et al.,[6] it remains stubbornly difficult to tap into this easily accessible information for a wider range of clinical outcomes. Even in the wake of a pandemic, where immense economical resources have been made available to develop a breath test to detect SARS-CoV-2, the use of human breath for clinical purposes remains a major challenge.

Secondary electrospray ionization (SESI[7]) combined with modern high resolution mass spectrometry (HRMS) has enabled over the last 15 years the real-time detection of several key metabolites with an unprecedented coverage.[8–12] Real-time analysis has the ability to minimize the risk of compromising the sample integrity by skipping sample pre-treatment and storage.[13–15] and to provide the result turnaround time in the order of minutes. Following years of technological development, optimization,[16] and elaboration of standardization procedures[17,18] we have recently integrated SESI-HRMS platforms in selected clinical settings and provided proof-of-principle for the use of this technique for the therapeutic management of patients with epilepsy.[19] Additionally, if needed SESI-HRMS can also be used for real-time quantification of breath metabolites.[11,20,21] However, the exploration of further clinical applications, and eventually regulatory approval, require large-scale multicenter studies and a better understanding on the core metabolic information captured in exhaled breath by SESI-HRMS, which remains largely unknown.

In this study, we aimed to set the basis for large-scale multicenter studies by implementing a previously demonstrated standard operating procedure (SOP) for data collection[17] and our patented data processing pipeline for breath analysis by SESI-HRMS (European patent No. 20186274.5 and 21185400.5).

[1]University Children's Hospital Basel UKBB, University of Basel, 4056 Basel, Switzerland

[2]Department of Biomedical Engineering, University of Basel, 4123 Allschwil, Switzerland

[3]Institute of Mass Spectrometry and Atmospheric Environment, Jinan University, Guangzhou 510632, China

[4]University Hospital Zurich USZ, 8091 Zurich, Switzerland

[5]These authors contributed equally

[6]Lead contact

*Correspondence: pablo.sinues@unibas.ch

https://doi.org/10.1016/j.isci.2022.105557

Furthermore, our goal was to determine the metabolic compounds and pathways represented in exhaled breath, rather than merely describing mass-to-charge features. We show that the technical variability is in line with well-established mass spectrometry-driven platforms and unveil the areas of the human metabolic network covered by this breath analysis technique.

RESULTS

Technical variability across sites

An analytical platform with low technical variability (i.e., the variability observed during the repeated measurement of the same sample) is the main prerequisite to detect subtle and true biological variability among different samples. However, so far there exists no system suitability test for breath analysis by SESI-HRMS. Therefore, we developed a "quality control" procedure, involving the acquisition of an eight compounds gas mixture of known concentration (Table S1) and a subsequent comparison of the gas-mix signal intensities over time using an inhouse MATLAB app (Figure S1). This quality control procedure was performed every day before the acquisition of breath sample in all three study sites (Basel, Zurich, and Guangzhou) (Figure 1A). We used well established multivariate control chart based on Hotelling's T^2 statistic[22] to compare signals of standard gas-mix compounds over time. Figure 1B shows the control chart for the standard gas-mix measurement acquired on the day of the last breath measurement in each site. In all sites the last T^2 value falls within the control limits, indicating similar performance (with limited technical variability) of the individual SESI-HRMS instruments in all three study sites during the course of the study period. The mean of the coefficients of variation (CV) of RAW signal intensities across all standard gas-mix compounds per site ranged from 18.4% to 22.1% (Table S2). Whereas, the mean CV of RAW signal intensities across all three sites per standard compound ranged from 18.5% to 21.7% (Table S2). Taken together, we claim to have an instrument setup with technical variability of ∼20% (assessed by the CV of standard gas-mix).

Multicentric real-time exhaled breath metabolomics

To determine the core composition of volatile organic compounds, present in the exhaled breath, we acquired 255 breath samples from nine healthy adults (30 ± 5 years, mean ± SD; four females and five males) in three study sites, using identical equipment and adhering to a data collection SOP. All breath samples were collected within six weeks, with one to two samples per working day (Figure 1C). After the completion of the measurements, breath data from all three study sites were analyzed using our patented data processing pipeline (European patent No. 20186274.5 and 21185400.5, for more details, see STAR methods).

In total, considering all nine subjects, 3198 features correlating with exhalations ($\rho_{spearman} > = 0.7$ and FDR= 0.01; i.e., breath-related features) were identified (2535 features in positive and 663 features in negative ionization mode). On average per subject, we detected ∼900 features in positive and ∼500 features in negative ionization mode (Figure S2). The majority of these features were detected within the *m/z* range of 70-300 (Figure S2). To remove sporadic features at the subject level, only features present in at least 80% of measurements from that subject were further considered. This resulted in a total of 1953 features. An overlap analysis of these features (Figure S3) revealed that 368 features appeared in all subjects. To further reduce the complexity, we looked at the overlap of these features at the site level rather than the subject level (Figure 1D). Hereafter, only 846 features occurring in at least two sites were considered core breath features and were used for further analysis. Despite adhering to a data collection SOP, as expected, we observed a batch (site) effect (Figure 1E). To correct this, we used the batch correction method ComBat[23] (Figure 1F). Finally, the resulting matrix (Figure 1G) of the ComBat corrected signal intensities of the 846 core features was used to identify the metabolic coverage of SESI-HRMS and to analyze biological variability.

Mapping mass spectral features into tangible metabolic information

Because of the real-time nature of SESI-HRMS, breath studies utilizing this technique often remain at the *m/z* level of information. Understanding which portion of the human metabolome is represented in the core breath signature is key to enable hypothesis-driven clinical studies. To address this issue, we used two different computational approaches to translate core breath *m/z* features into compounds. These approaches rely on the accurate mass obtained from the Orbitrap instrument (typically within a tolerance of ±2 ppm).[24] In the first approach, breath features were arranged based on the mass differences into a map, where each node is a *m/z* feature assigned to a molecular formula, to identify homologous series in the C, H and O space (Data S1, Figure 2A is a subset of Data S1). We employ previously confirmed feature-to-compound assignment at levels 1 and 2[25] for saturated fatty acids,[8] amino acids,[11] aldehydes,[9]

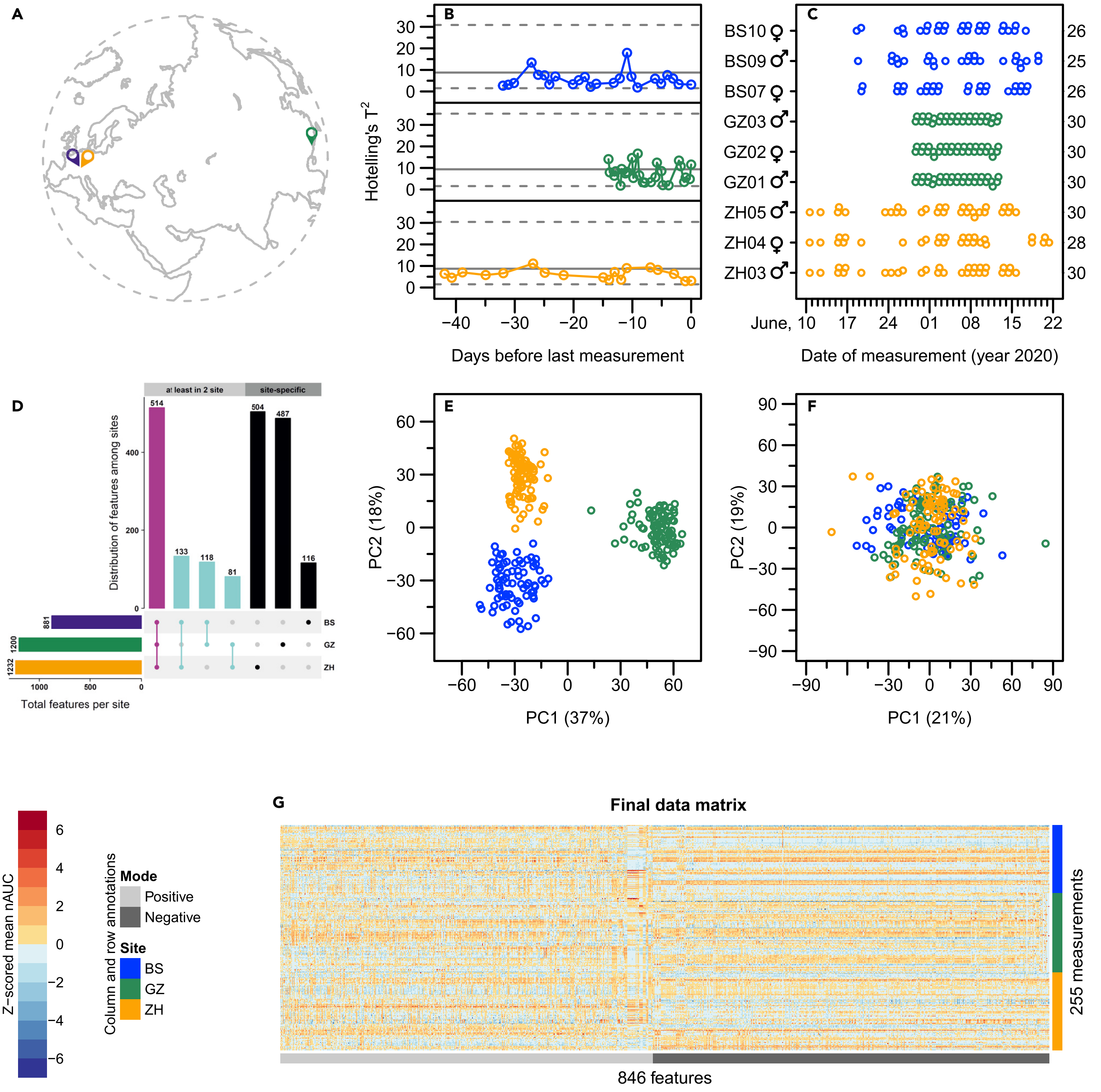


Figure 1. Overview of the multicentric study

(A) This multicenter study was conducted in three sites (Basel in blue, Zurich in orange, and Guangzhou in green).

(B) Hotelling's T^2 control chart for standard gas-mix compounds acquired on the last day of measurement at each site.

(C) Overview of study duration and number of included breath measurements per subject (number on the right side). In total, we analyzed 255 breath samples.

(D) Upset plot representing the overlap of breath features across sites.

(E) PCA plot of 846 core breath features (present in at least two sites) before batch correction.

(F) PCA plot of core breath data after batch correction using ComBat.

(G) Finally, a ComBat corrected 255 × 846 (samples × features) data matrix (is only used here to ease visual representation; actual downstream analysis was done on raw numbers) was obtained and used further for the analysis of biological variability and metabolic coverage.

nitro-aromatic compounds,[26] and tricarboxylic acid (TCA) cycle compounds,[10] as anchor points to infer the chemical formulas of the neighboring features. For example, the grid containing 4-hydroxy-2-alkenals suggests that further series of aldehydes with up to six double bond equivalents (DBE) are detectable in exhaled breath. Similarly, the grid comprising saturated fatty acids expanding from acetic acid to palmitic acid, suggests that closely related unsaturated fatty acids (up to four DBE) may also be present in breath.

We hypothesized that metabolites which are close in the chemical space, are also closely related in the metabolic/biochemical space. To test this, we computed pairwise correlations between signal intensities of the 846 core breath features (Figure 2B). This network architecture suggests that features with similar chemical formulas tend to cluster together. Closer inspection reveals fine structures that resemble what one would expect from a metabolic perspective. For example, short-chain (C_1-C_4) fatty acids cluster together, separated from the rest of network, perhaps due to the fact that these are byproducts of gut microbiome.[27] A distinct tail is formed by long-chain (C_{14}-C_{18}) fatty acids, whereas medium-chain (C_5-C_{13}) fatty acid blend in the main structure of the correlation network. Similarly, TCA cycle compounds cluster together and close to a feature corresponding to a simple sugar with formula $C_6H_{12}O_6$. Potential oxidative stress markers such as nitrophenols also naturally cluster together. Additionally, a gradient of color could be observed in the network clusters, referring to compounds with an increasing (decreasing) oxygenation level across the correlation clusters.

In the second approach, MetaboAnalystR (version 3.0.3)[28] was used to assign breath features to the main lipid and non-lipid class of compounds from the RefMet database[29] considering only protonated and deprotonated forms within a tolerance of 2 ppm. Assigned compounds were then cross-referenced to other common databases such as PubChem, HMDB, KEGG, and METLIN (Figure 3A). As a result, we were able to assign 846 core breath features to 477 unique KEGG compounds. These compounds mainly belong to amino acids, xenobiotics biodegradation, and carbohydrate metabolic pathways from the KEGG's "Metabolism" pathway class (Figure 3B). Additionally, we calculated intra-subject CVs and performed a one-way ANOVA to examine the variability in the signal intensities of core breath features between the nine subjects (Data S2). Feature level p-values from the ANOVA were then combined as needed to estimate compound level p-values (Fisher's method). Later these combined p-values were used to calculate the proportion of compounds that varied significantly (combined $p\text{-value} < 0.05$) between subjects in the different pathways (red bars in Figure 3B). This analysis revealed a significant inter-subject variability for ~50-65% of the compounds of the top-three pathways. Notable exceptions for regions of high stability (i.e. no significant differences across subjects) include the TCA cycle (Figure 3C), despite the fact that ~60% of the carbohydrate metabolic pathway compounds were significantly different between subjects (Figure 3B). Subsequently, we created a simplified custom map considering few compounds from the most represented KEGG pathways (Figure 4A) and as expected, features corresponding to compounds from the same pathway clustered together in the correlation network (Figure 4B). Furthermore, to provide even more comprehensive coverage of feature to molecular formula assignment, we queried all *m/z* breath features against all possible molecular formulae (with the combination of the following elements: C, H, N, and O, including also ^{13}C isotopic form) under 1000 Da following the "seven golden rules," as described by Kind and Fiehn,[30] again considering only (de)protonated species within a tolerance of 2 ppm (Data S2). To gain more confidence in these assignments, we compared the observed isotopic distribution of a feature and the theoretical isotopic distribution of the assigned formula, and annotated each assignment as pass or fail "isotope filter" (to pass the isotope filter, first three isotopic peaks from observed and theoretical distributions must match with an overall intensity correlation higher than 0.8).

DISCUSSION

Clinical biomarker discovery is a key element in the progression towards precision medicine.[31] Coordinated multicentric studies involving the analysis of thousands of samples remain at the heart of such efforts.[32] Due to its non-invasiveness and readily availability, exhaled breath appears to be a perfect candidate for such high throughput endeavors. However, the adoption of breath metabolomics as a widespread branch of the multiple -omics platforms requires a "big science" approach.[33] Indeed, the standardization of data collection and data processing have been highlighted as main priorities in the field of breath research.[34,35] The adoption of a common analytical platform across multiple centers following carefully designed and optimized SOPs will unlock this potential, overcoming the often-patchy breath metabolomics studies, whereby the use of different analysis techniques and small samples sizes leads to little progress in the search for exhaled biomarkers. In this work, we build upon fifteen years of experience developing SESI-HRMS for breath metabolomics to deploy for the first time an interoperability framework.

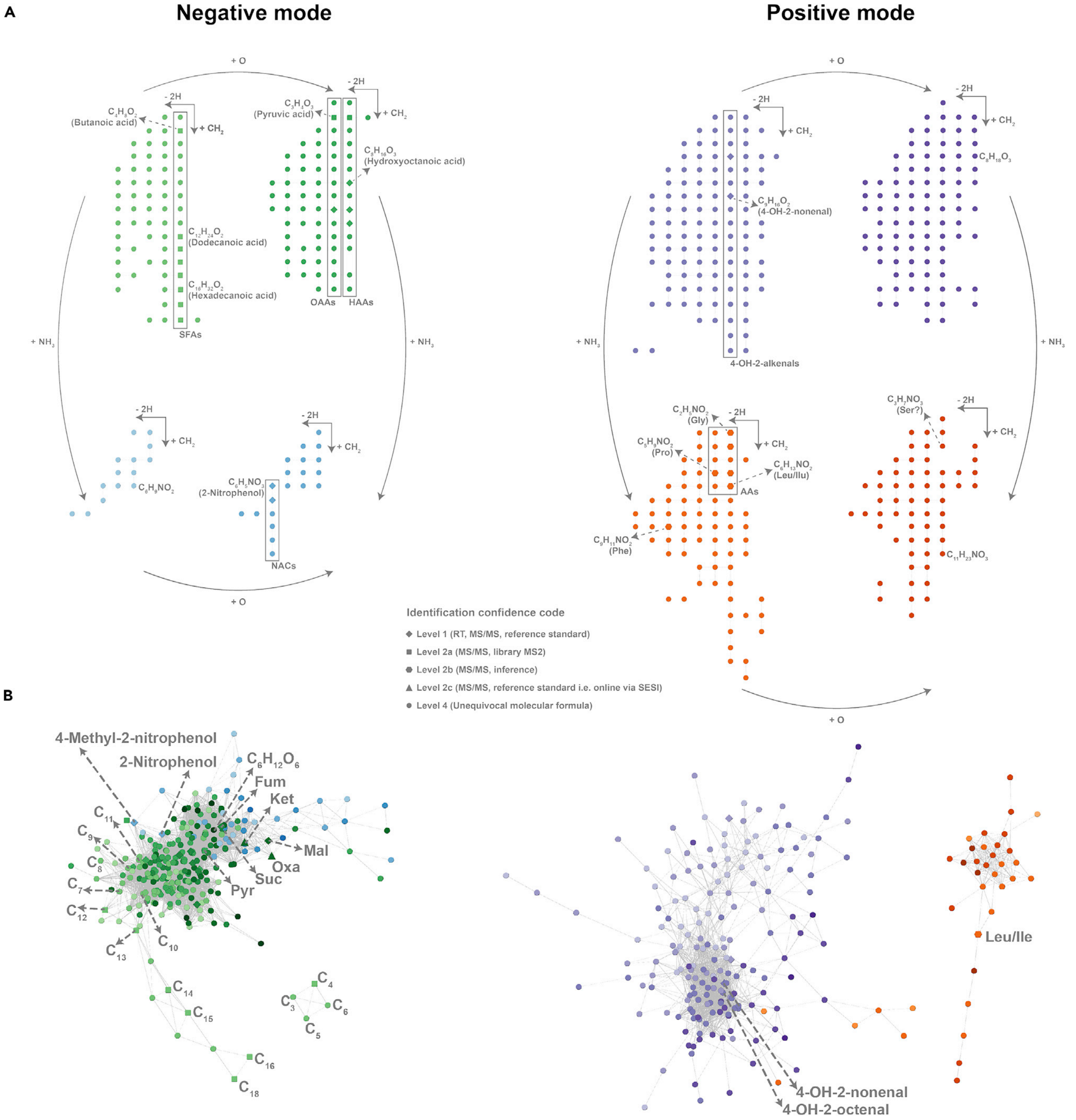

A
Negative mode
Positive mode
+ O
- 2H
+ CH_2
$C_4H_8O_2$ (Butanoic acid)
$C_3H_4O_3$ (Pyruvic acid)
$C_8H_{16}O_3$ (Hydroxyoctanoic acid)
$C_{12}H_{24}O_2$ (Dodecanoic acid)
$C_{16}H_{32}O_2$ (Hexadecanoic acid)
SFAs
OAAs
HAAs
+ NH_3
$C_6H_9NO_2$
$C_6H_5NO_3$ (2-Nitrophenol)
NACs
$C_9H_{16}O_2$ (4-OH-2-nonenal)
4-OH-2-alkenals
$C_8H_{18}O_3$
$C_2H_5NO_2$ (Gly)
$C_5H_9NO_2$ (Pro)
$C_6H_{13}NO_2$ (Leu/Ilu)
AAs
$C_9H_{11}NO_2$ (Phe)
$C_3H_7NO_3$ (Ser?)
$C_{11}H_{23}NO_3$
Identification confidence code
◆ Level 1 (RT, MS/MS, reference standard)
■ Level 2a (MS/MS, library MS2)
⬢ Level 2b (MS/MS, inference)
▲ Level 2c (MS/MS, reference standard i.e. online via SESI)
● Level 4 (Unequivocal molecular formula)
B
4-Methyl-2-nitrophenol
2-Nitrophenol
$C_6H_{12}O_6$
Fum
Ket
Mal
Oxa
Suc
Pyr
C_{11}
C_9
C_8
C_7
C_{12}
C_{13}
C_{10}
C_{14}
C_{15}
C_{16}
C_{18}
C_4
C_3
C_6
C_5
Leu/Ile
4-OH-2-nonenal
4-OH-2-octenal

Figure 2. Mapping the chemical space

(A) Each node is a measured mass spectral feature arranged based on mass difference. Specifically, within one cluster, mass increase in the multiple of 14.01565 Da (corresponding to CH_2) from top to bottom and decrease in the multiple of 2.01565 Da (corresponding to 2H, i.e., formation of double bound) from right to left. Additionally, clusters horizontally differ by 15.99492 Da (corresponding to O) and vertically differ by 17.02655 Da (corresponding to NH_3). Some of these features were previously assigned to compounds with various degrees of confidence, this is reflected by node shape. Some of these assignments are annotated in the figure to ease interpretation. Nodes are colored based on estimated formulas (negative mode: non-nitrogen compounds in green and nitrogen-containing compounds in blue; positive mode: non-nitrogen compound in purple and nitrogen-containing compounds in orange). Darker shades indicate compounds with a greater number of oxygen atoms. See Cytoscape file in Data S1 for more features and detailed annotation.

(B) Correlation network ($\rho_{spearman} > = 0.8$ and p-value $< = 0.01$) of 846 core breath features suggests features with similar chemical formulas tend to cluster together. Node shape and color follow the same rule as in panel A.

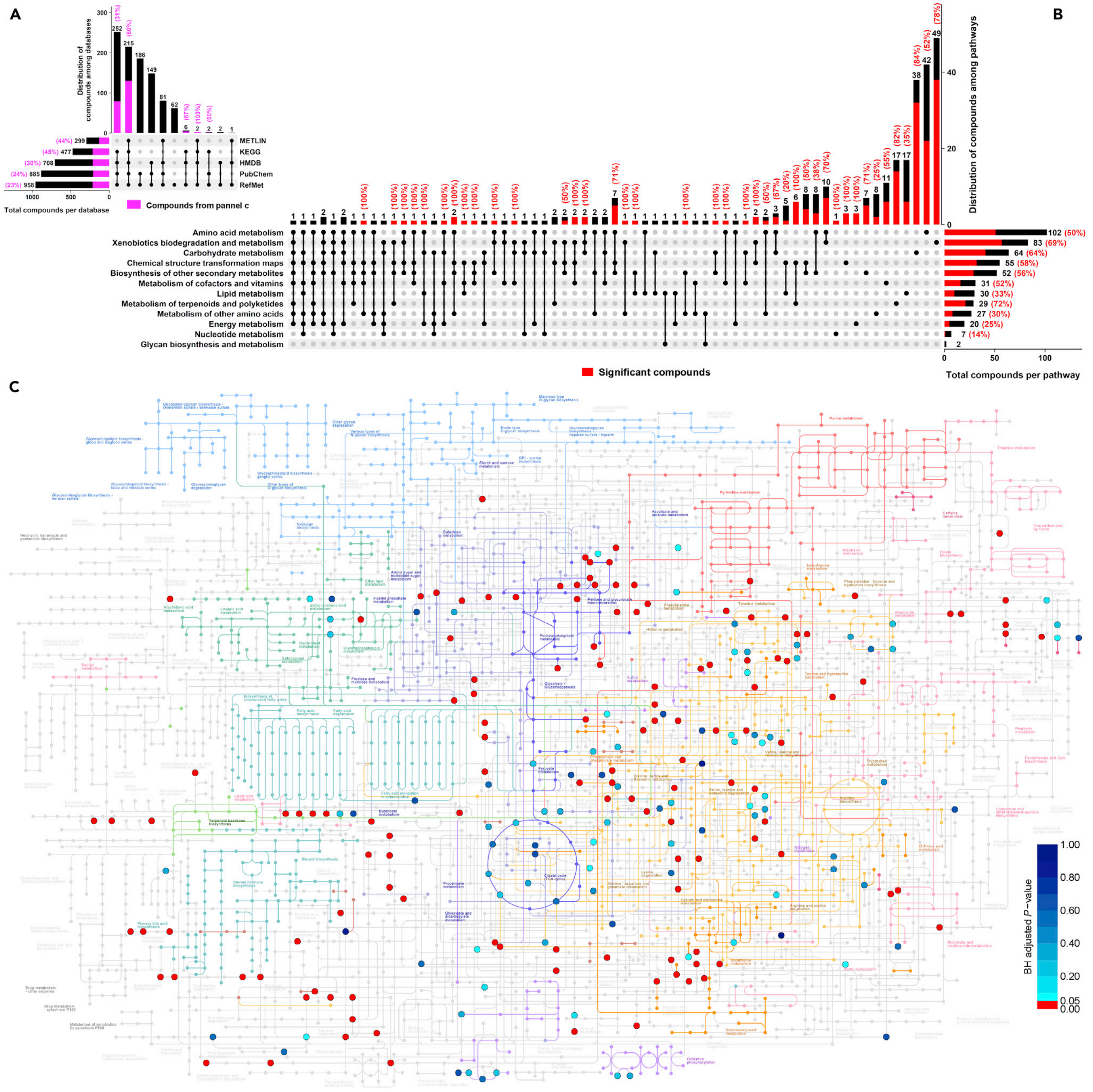


Figure 3. Mapping of the metabolic network and inter-subject stability regions

(A) Upset plot representing the overlap of compounds assigned to core breath features from different databases. Magenta overlay shows the number of compounds from KEGG map 01100.

(B) Upset plot indicating the overlap of compounds from KEGG across various pathways from "Metabolism" pathway class. Red overlay bars show the proportion of compounds with significantly different levels among subjects.

(C) Figure shows the KEGG's metabolic pathways map (i.e., map 01100), where compounds corresponding to core breath features are emphasized by bigger dots. Color of dot is based on BH-adjusted p-values from ANOVA across subjects at the feature level.

The use of an external standard gas-mix led us to conclude that the technical variability of a SESI-HRMS instrument is around 20%, which is in line with the variability often reported for mass spectrometry-based metabolomics analyses.[36] Besides standardized data collection, a standardized procedure for data processing is also key to obtain comparable results.[34] Because the SESI-HRMS data structure lacks chromatographic separation, the most common algorithms for pre-processing metabolomics data (e.g. XCMS[37]) are not suitable for this

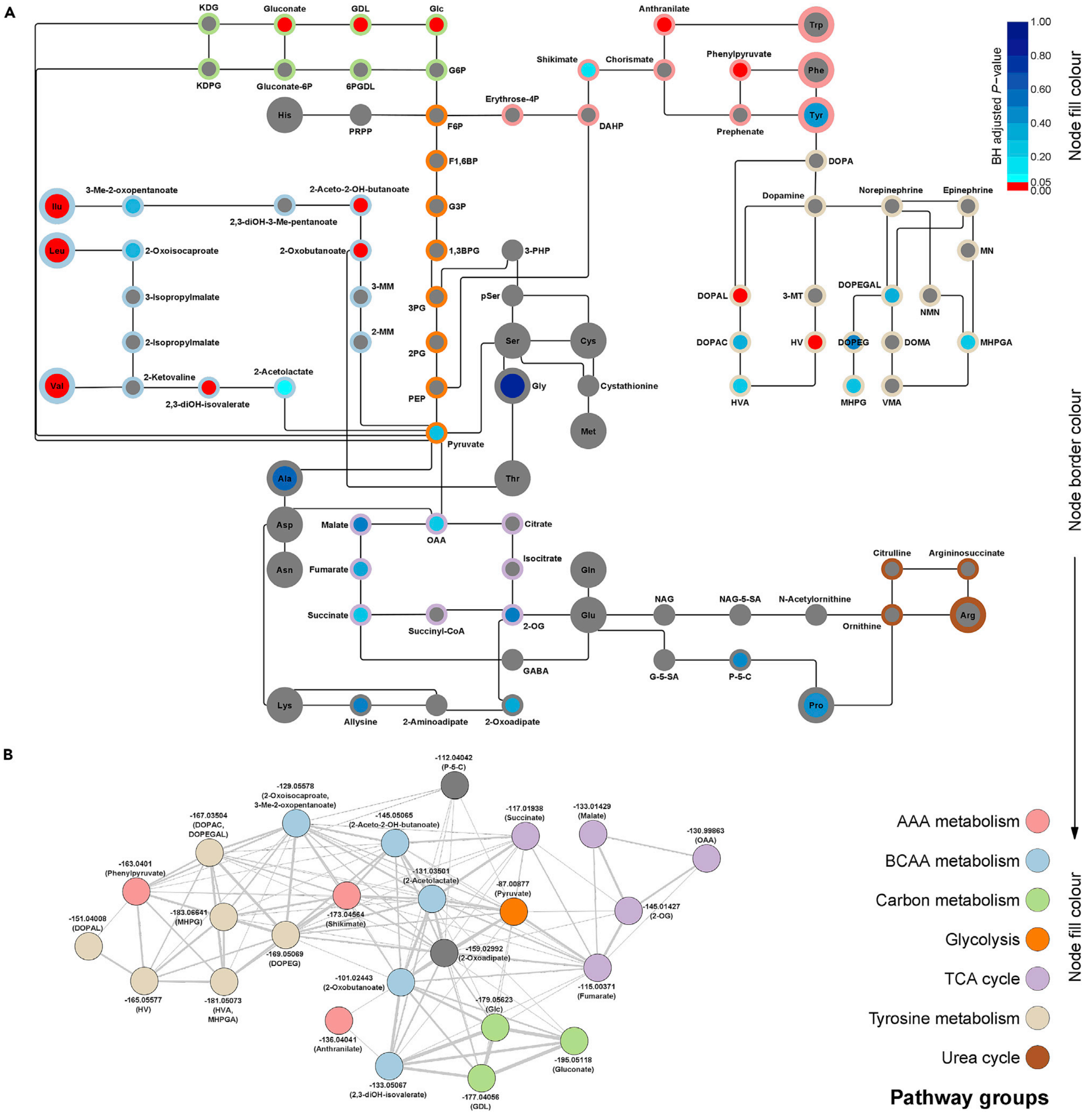


Figure 4. Features corresponding to compounds from the same pathway tend to cluster together in the correlation network

(A) A simplified custom map (based on KEGG's map 01100) considering the most represented pathways of exhaled breath. Each node represents a compound and is colored based on p-value from ANOVA across subjects. Gray nodes represent compounds not detected in core breath features. Whole map is also divided into different pathway groups, as represented by node border color.

(B) Correlation network ($\rho_{spearman}$ > = 0.8 and p-value < = 0.01) of features corresponding to compounds from panel A, where node color represents the pathway group and edge line width represents the strength of correlation. Gray nodes are not part of any pathway group.

purpose. Recently, we developed a pre-processing pipeline based on the parallel collection of exhaled CO_2 to ensure the downstream analysis of the exhaled end-tidal fraction (European patents No. 20186274.5 and 21185400.5).[17] As previously reported,[17] we observed subject-independent but molecule-dependent exhalation patterns for selected aldehyde series for all subjects (Figure S4), highlighting the reproducibility of SESI-HRMS system. Furthermore, compared to the CV of standard gas-mix (note the use of fifth root transformed values) we observed higher intra-subject CVs for core breath features in all subjects (Figure S5), indicating SESI-HRMS system is suitable to capture inherent biological variability of exhaled breath. Despite highly standardized data collection procedures, a clear site-driven batch effect was observed. Such an issue is a typical problem of, not only metabolomics-driven mass spectrometry, but rather a problem afflicting all almost -omics data and platforms.[38] As a matter of fact, we borrowed the batch correction method ComBat, which was originally developed to adjust microarray expression datasets.[39] To minimize the risk of detecting false significant results, no covariate other than batch (which clearly represent different sites and is low in number compared to samples) was used during ComBat correction.[40,41]

Once we ensured the highest possible data quality, the next question we addressed was which metabolic information is encoded in the core breath fingerprint. During the last years, a number of attempts have been carried out to identify the chemical structure of a handful of detected metabolites (e.g. ω-oxidation pathway of fatty acids[8,12,19,26] and multiple amino acid pathways[11,19,42]). However, this bottom-up approach seems largely impractical to cover the ~850 breath features detected here. For this reason, we propose a top-down computational approach combining mass difference to map the chemical space (Figure 2A), correlation networks (Figures 2B and 4B) to verify whether they resemble biochemical pathways (Figure 4A) and querying public databases (Figure 3). This strategy unveiled that most of the detected metabolites can be mapped to the amino acid, xenobiotic, and carbohydrate metabolic pathways. This is consistent with previous clinical studies suggesting that indeed some of these pathways (i.e. amino acid metabolism) are altered in patients with epilepsy.[19] Interestingly, we found some metabolic pathways that varied significantly between subjects, whereas certain specific regions (e.g. TCA cycle) varied little. Overall, around ~50% of the detected metabolites varied little between subjects. This insight can be critical for future clinical studies using breath analysis.

In conclusion, in this study, we demonstrated an interoperability framework for standardized and harmonized data collection and data processing for breath analysis with SESI-HRMS. Using two different computational approaches, we mapped the biochemical space covered by the main metabolites detected by this technique. We unveiled previously undetected homologous series of aldehydes and fatty acids in breath and found amino acid, xenobiotic, and carbohydrate metabolic pathways to be the most covered ones. The interoperability framework and the insights into the biological variability and metabolic coverage presented here will help to design and conduct large-scale multicentric clinical studies in precision medicine using breath metabolomics. Future hypothesis-driven clinical studies may target conditions where alterations of the above-mentioned pathways are expected.

Limitations of the study

This study has some limitations to be noted. First, the number of sites and subjects included was rather small, possibly leading to an overestimation of biological variability. Second, the identification of metabolites in breath was based on database matching using the measured accurate masses (within 2 ppm), therefore biochemical interpretation should be done with caution. Further, UPLC-MS/MS analysis using chemical standards would be required to increase the confidence in compound identification. Third, we did not have detailed data on the subjects' diet or other factors that might influence metabolism, therefore we could not further assess the inter-subject variability attributable to these factors. However, the herein presented results are of sufficient precision to guide future large-scale multicentric clinic trials.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Untargeted pre-processing of mass spectral data

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.105557.

ACKNOWLEDGMENTS

We thank the staff from the collaborating sites in Basel, Zurich, and Guangzhou for their contribution to this study. This work is part of the Zurich Exhalomics project under the umbrella of University Medicine Zurich/Hochschulmedizin Zürich. This study was funded by the Fondation Botnar, Switzerland (Professorship of Pablo Sinues) and the Swiss National Science Foundation (grant number 320030_173168 and PCEGP3_181300).

AUTHOR CONTRIBUTIONS

P.S., X.L., M.K., and U.F. designed the study. A.G., M.O., J.Z., M.A., F.D., F.S., X.C., X.L., and N.S. were responsible for the data collection. A.G. and K.S. performed the data analysis and wrote the main article with input from the co-authors. P.S., U.F., M.K., J.U., and X.L. contributed to the interpretation of results. All authors read and approved the final article.

DECLARATION OF INTERESTS

P.S. and M.K. are co-founders of Deep Breath Intelligence AG (Switzerland). K.S. and F.S. are consultants for Deep Breath Intelligence AG (Switzerland). P.S., K.S., A.G., M.K., U.F., and M.O. hold a patent on data processing for breath analysis by SESI-HRMS. P.S. and K.S. hold a patent on a system suitability test for SESI-HRMS. All other authors declare no competing interests.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: April 3, 2022
Revised: September 10, 2022
Accepted: November 9, 2022
Published: December 22, 2022

REFERENCES

1. Nicholson, J.K., Holmes, E., Kinross, J.M., Darzi, A.W., Takats, Z., and Lindon, J.C. (2012). Metabolic phenotyping in clinical and surgical environments. Nature *491*, 384–392. https://doi.org/10.1038/nature11708.

2. Mirnezami, R., Nicholson, J., and Darzi, A. (2012). Preparing for precision medicine. N. Engl. J. Med. *366*, 489–491. https://doi.org/10.1056/NEJMp1114866.

3. Massaro, A.F., Gaston, B., Kita, D., Fanta, C., Stamler, J.S., and Drazen, J.M. (1995). Expired nitric oxide levels during treatment of acute asthma. Am. J. Respir. Crit. Care Med. *152*, 800–803. https://doi.org/10.1164/ajrccm.152.2.7633745.

4. Holguin, F., Cardet, J.C., Chung, K.F., Diver, S., Ferreira, D.S., Fitzpatrick, A., Gaga, M., Kellermeyer, L., Khurana, S., Knight, S., et al. (2020). Management of severe asthma: a European respiratory society/American thoracic society guideline. Eur. Respir. J. *55*, 1900588. https://doi.org/10.1183/13993003.00588-2019.

5. Kato, M., Saito, M., Fukuda, S., Kato, C., Ohara, S., Hamada, S., Nagashima, R., Obara, K., Suzuki, M., Honda, H., et al. (2004). 13C-Urea breath test, using a new compact nondispersive isotope-selective infrared spectrophotometer: comparison with mass spectrometry. J. Gastroenterol. *39*, 629–634.

6. Pauling, L., Robinson, A.B., Teranishi, R., and Cary, P. (1971). Quantitative analysis of urine vapor and breath by gas-liquid partition chromatography. Proc. Natl. Acad. Sci. USA *68*, 2374–2376. https://doi.org/10.1073/pnas.68.10.2374.

7. Whitehouse, C.M., L., F., Meng, C.K., and Fenn, J.B. (1986). Proceedings of the 34th ASMS Conference on Mass Spectrometry and Allied Topics (ASMS Conference on Mass Spectrometry and Allied Topics).

8. Martínez-Lozano, P., and Fernández de la Mora, J. (2008). Direct analysis of fatty acid vapors in breath by electrospray ionization and atmospheric pressure ionization-mass spectrometry. Anal. Chem. *80*, 8210–8215. https://doi.org/10.1021/ac801185e.

9. García-Gómez, D., Martínez-Lozano Sinues, P., Barrios-Collado, C., Vidal-de-Miguel, G., Gaugg, M., and Zenobi, R. (2015). Identification of 2-alkenals, 4-Hydroxy-2-alkenals, and 4-hydroxy-2, 6-alkadienals in exhaled breath condensate by UHPLC-HRMS and in breath by real-time HRMS. Anal. Chem. *87*, 3087–3093. https://doi.org/10.1021/ac504796p.

10. Tejero Rioseras, A., Singh, K.D., Nowak, N., Gaugg, M.T., Bruderer, T., Zenobi, R., and Sinues, P.M.L. (2018). Real-time monitoring of tricarboxylic acid metabolites in exhaled breath. Anal. Chem. *90*, 6453–6460. https://doi.org/10.1021/acs.analchem.7b04600.

11. García-Gómez, D., Gaisl, T., Bregy, L., Cremonesi, A., Sinues, P.M.L., Kohler, M., and Zenobi, R. (2016). Real-time quantification of amino acids in the exhalome by secondary electrospray ionization-mass spectrometry: a proof-of-principle study. Clin. Chem. *62*, 1230–1237. https://doi.org/10.1373/clinchem.2016.256909.

12. Gaugg, M.T., Bruderer, T., Nowak, N., Eiffert, L., Martinez-Lozano Sinues, P., Kohler, M., and Zenobi, R. (2017). Mass-spectrometric detection of omega-oxidation products of aliphatic fatty acids in exhaled breath. Anal. Chem. *89*, 10329–10334. https://doi.org/10.1021/acs.analchem.7b02092.

13. McLerran, D., Grizzle, W.E., Feng, Z., Thompson, I.M., Bigbee, W.L., Cazares, L.H., Chan, D.W., Dahlgren, J., Diaz, J., Kagan, J., et al. (2008). SELDI-TOF MS whole serum proteomic profiling with IMAC surface does not reliably detect prostate cancer. Clin. Chem. *54*, 53–60. https://doi.org/10.1373/clinchem.2007.091496.

14. McLerran, D., Grizzle, W.E., Feng, Z., Bigbee, W.L., Banez, L.L., Cazares, L.H., Chan, D.W., Diaz, J., Izbicka, E., Kagan, J., et al. (2008). Analytical validation of serum proteomic profiling for diagnosis of prostate cancer: sources of sample bias. Clin. Chem. *54*, 44–52. https://doi.org/10.1373/clinchem.2007.091470.

15. Ghini, V., Abuja, P.M., Polasek, O., Kozera, L., Laiho, P., Anton, G., Zins, M., Klovins, J., Metspalu, A., Wichmann, H.E., et al. (2021). Metabolomic fingerprints in large population cohorts: impact of preanalytical heterogeneity. Clin. Chem. *67*, 1153–1155. https://doi.org/10.1093/clinchem/hvab092.

16. Barrios-Collado, C., Vidal-de-Miguel, G., and Martinez-Lozano Sinues, P. (2016). Numerical modeling and experimental validation of a universal secondary electrospray ionization source for mass spectrometric gas analysis in real-time. Sensor. Actuator. B Chem. *223*, 217–225. https://doi.org/10.1016/j.snb.2015.09.073.

17. Singh, K.D., Tancev, G., Decrue, F., Usemann, J., Appenzeller, R., Barreiro, P., Jaumà, G., Macia Santiago, M., Vidal de Miguel, G., Frey, U., and Sinues, P. (2019). Standardization procedures for real-time breath analysis by secondary electrospray ionization high-resolution mass spectrometry. Anal. Bioanal. Chem. *411*, 4883–4898. https://doi.org/10.1007/s00216-019-01764-8.

18. Gisler, A., Lan, J., Singh, K.D., Usemann, J., Frey, U., Zenobi, R., and Sinues, P. (2020). Real-time breath analysis of exhaled compounds upon peppermint oil ingestion by secondary electrospray ionization-high resolution mass spectrometry: technical aspects. J. Breath Res. *14*, 046001. https://doi.org/10.1088/1752-7163/ab9f8b.

19. Singh, K.D., Osswald, M., Ziesenitz, V.C., Awchi, M., Usemann, J., Imbach, L.L., Kohler, M., García-Gómez, D., van den Anker, J., Frey, U., et al. (2021). Personalised therapeutic management of epileptic patients guided by pathway-driven breath metabolomics. Commun. Med. *1*, 21. https://doi.org/10.1038/s43856-021-00021-3.

20. Liu, C., Zeng, J., Sinues, P., Fang, M., Zhou, Z., and Li, X. (2021). Quantification of volatile organic compounds by secondary electrospray ionization-high resolution mass spectrometry. Anal. Chim. Acta *1180*, 338876. https://doi.org/10.1016/j.aca.2021.338876.

21. Martínez-Lozano, P., Zingaro, L., Finiguerra, A., and Cristoni, S. (2011). Secondary electrospray ionization-mass spectrometry: breath study on a control group. J. Breath Res. *5*, 016002. https://doi.org/10.1088/1752-7155/5/1/016002.

22. MacGregor, J.F., and Kourti, T. (1995). Statistical process control of multivariate processes. Control Eng. Pract. *3*, 403–414. https://doi.org/10.1016/0967-0661(95)00014-L.

23. Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E., and Storey, J.D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics *28*, 882–883. https://doi.org/10.1093/bioinformatics/bts034.

24. Strupat, K., Scheibner, O., and Bromirski, M. (2013). High-resolution, Accurate-Mass Orbitrap Mass Spectrometry–Definitions, Opportunities, and Advantages (Thermo Fisher Scientific (Bremen) GmbH).

25. Schymanski, E.L., Jeon, J., Gulde, R., Fenner, K., Ruff, M., Singer, H.P., and Hollender, J. (2014). Identifying small molecules via high resolution mass spectrometry: communicating confidence. Environ. Sci. Technol. *48*, 2097–2098. https://doi.org/10.1021/es5002105.

26. Gaugg, M.T., Nussbaumer-Ochsner, Y., Bregy, L., Engler, A., Stebler, N., Gaisl, T., Bruderer, T., Nowak, N., Sinues, P., Zenobi, R., and Kohler, M. (2019). Real-time breath analysis reveals specific metabolic signatures of COPD exacerbations. Chest *156*, 269–276. https://doi.org/10.1016/j.chest.2018.12.023.

27. Koh, A., De Vadder, F., Kovatcheva-Datchary, P., and Bäckhed, F. (2016). From dietary fiber to host physiology: short-chain fatty acids as key bacterial metabolites. Cell *165*, 1332–1345. https://doi.org/10.1016/j.cell.2016.05.041.

28. Chong, J., and Xia, J. (2018). MetaboAnalystR: an R package for flexible and reproducible analysis of metabolomics data. Bioinformatics *34*, 4313–4314. https://doi.org/10.1093/bioinformatics/bty528.

29. Fahy, E., and Subramaniam, S. (2020). RefMet: a reference nomenclature for metabolomics. Nat. Methods *17*, 1173–1174. https://doi.org/10.1038/s41592-020-01009-y.

30. Kind, T., and Fiehn, O. (2007). Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. BMC Bioinf. *8*, 105. https://doi.org/10.1186/1471-2105-8-105.

31. Neumann, S. (2015). Biomarkers – past and future. In Biomarker Validation (Wiley-VCH Verlag GmbH & Co. KGaA), pp. 1–22. https://doi.org/10.1002/9783527680658.ch1.

32. Kachroo, P., Stewart, I.D., Kelly, R.S., Stav, M., Mendez, K., Dahlin, A., Soeteman, D.I., Chu, S.H., Huang, M., Cote, M., et al. (2022). Metabolomic profiling reveals extensive adrenal suppression due to inhaled corticosteroid therapy in asthma. Nat. Med. *28*, 814–822. https://doi.org/10.1038/s41591-022-01714-5.

33. Poste, G. (2011). Bring on the biomarkers. Nature *469*, 156–157. https://doi.org/10.1038/469156a.

34. Horváth, I., Barnes, P.J., Loukides, S., Sterk, P.J., Högman, M., Olin, A.C., Amann, A., Antus, B., Baraldi, E., Bikov, A., et al. (2017). A European Respiratory Society technical standard: exhaled biomarkers in lung disease. Eur. Respir. J. *49*, 1600965. https://doi.org/10.1183/13993003.00965-2016.

35. Henderson, B., Ruszkiewicz, D.M., Wilkinson, M., Beauchamp, J.D., Cristescu, S.M., Fowler, S.J., Salman, D., Francesco, F.D., Koppen, G., Langejürgen, J., et al. (2020). A benchmarking protocol for breath analysis: the peppermint experiment. J. Breath Res. *14*, 046008. https://doi.org/10.1088/1752-7163/aba130.

36. Sampson, J.N., Boca, S.M., Shu, X.O., Stolzenberg-Solomon, R.Z., Matthews, C.E., Hsing, A.W., Tan, Y.T., Ji, B.T., Chow, W.H., Cai, Q., et al. (2013). Metabolomics in epidemiology: sources of variability in metabolite measurements and implications. Cancer Epidemiol. Biomarkers Prev. *22*, 631–640. https://doi.org/10.1158/1055-9965.EPI-12-1109.

37. Smith, C.A., Want, E.J., O'Maille, G., Abagyan, R., and Siuzdak, G. (2006). XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Anal. Chem. *78*, 779–787. https://doi.org/10.1021/ac051437y.

38. Goh, W.W.B., Wang, W., and Wong, L. (2017). Why batch effects matter in omics data, and how to avoid them. Trends Biotechnol. *35*, 498–507. https://doi.org/10.1016/j.tibtech.2017.02.012.

39. Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics *8*, 118–127. https://doi.org/10.1093/biostatistics/kxj037.

40. Zindler, T., Frieling, H., Neyazi, A., Bleich, S., and Friedel, E. (2020). Simulating ComBat: how batch correction can lead to the systematic introduction of false positive results in DNA methylation microarray studies. BMC Bioinf. *21*, 271. https://doi.org/10.1186/s12859-020-03559-6.

41. Nygaard, V., Rødland, E.A., and Hovig, E. (2016). Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. Biostatistics *17*, 29–39. https://doi.org/10.1093/biostatistics/kxv027.

42. García-Gómez, D., Gaisl, T., Bregy, L., Martínez-Lozano Sinues, P., Kohler, M., and Zenobi, R. (2016). Secondary electrospray ionization coupled to high-resolution mass spectrometry reveals tryptophan pathway metabolites in exhaled human breath. Chem. Commun. *52*, 8526–8528. https://doi.org/10.1039/C6CC03070J.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Raw data	MetaboLights https://www.ebi.ac.uk/metabolights	MetaboLights: MTBLS5657
Software and algorithms		
MATLAB version 2021a	MathWorks Inc., USA	https://ch.mathworks.com
Thermo Exactive Plus Tune software version 2.9	Thermo Fisher Scientific, Germany	https://www.thermofisher.com
RawFileReader version 5.0.0.38	Thermo Fisher Scientific, Germany	https://www.thermofisher.com
Cytoscape version 3.8.2	Cytoscape	https://cytoscape.org, RRID: SCR_003032
MetaboAnalystR version 3.0.3	MetaboAnalyst	https://www.metaboanalyst.ca/, RRID: SCR_015539
R version 4.1.0	The Comprehensive R Archive Network (CRAN)	https://cran.r-project.org, RRID: SCR_001905
Other		
TaperTip silica capillary emitter	New Objective, USA	https://www.mswil.com
Bacterial filters	Vyaire Medical, USA	MicroGardTM II, https://www.vyaire.com
SESI source	FIT, Spain	SUPER SESI, https://www.fossiliontech.com
Mass spectrometer	Thermo Fisher Scientific, Germany	Q Exactive Plus, https://www.thermofisher.com
Gas dilution calibrator	Sabio Environmental, USA	Model 2010, https://www.sabio.com
Standard gas mixture	Dalian Special Gases, China	http://www.dl-gas.com

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact Pablo Sinues (pablo.sinues@unibas.ch).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The RAW files of the real-time breath measurements are available from the repository MetaboLights: MTBLS5657 (https://www.ebi.ac.uk/metabolights).
- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human breath samples

270 real-time breath measurements were acquired from nine healthy adults (30 ± 5 years, mean ± SD; four females and five males; 30 measurements each) in three sites (Zurich and Basel in Switzerland and Guangzhou in China) following data acquisition SOP. Due to deviation from the specified SOP, 15 breath measurements were excluded from further analysis, leading to a total of 255 measurements. Each measurement consists of 5–6 prolonged exhalations directly into the SESI-HRMS system both in positive and negative

ion mode (resulting in 510 RAW data files). Typically, each measurement takes 15–20 min to finish and subject performed 1–2 measurements per day during the time frame of six weeks.

Subjects refrained from using cosmetics, eating or drinking anything (except water) and cleaning their mouth (i.e. brush, mouthwash etc.) for at least 1 h prior to the breath measurement in order to avoid confounding mass features. The Ethics Committee of Northwest and Central Switzerland (ID 2018–01324) approved the study protocol and written informed consent was obtained at enrollment.

METHOD DETAILS

Analytical platform for breath samples

The SESI-HRMS analytical platform consisted of an exhalation interface (Exhalion, FIT Spain) for real-time time display of CO_2, flow rate, exhaled volume and pressure drop, and an ion source (Super SESI, FIT Spain) coupled to a high-resolution mass spectrometer (Q-Exactive Plus, Thermo Fisher Scientific, Germany). Exhalion was calibrated weekly with 5% CO_2. Commercially available bacterial filters (MicroGardTM II, Vyaire Medical, USA) were used as mouthpieces and replaced weekly. Mass spectra were acquired in full scan mode over a range from *m/z* 70–1000 with a resolution on the order of 140,000 for positive and negative modes. Two microscans, automatic gain control (AGC) target of 1e6 and maximum injection time of 500 ms were applied. Q Exactive Tune software (version 2.9) was used to directly control mass spectrometry for real-time measurements. Furthermore, the mass spectrometer was externally calibrated on a weekly basis using a customized calibration option and internally calibrated during each measurement by enabling lock mass (Table S3). For the electrospray formation, a 20-μm ID non-coated TaperTip silica capillary emitter (New Objective, Woburn, MA) and 0.1% formic acid in water were used with the following settings: sheath gas flow rate 60, auxiliary gas flow rate 2, spray voltage 3.5 kV, capillary temperature 275°C, and S-lens RF level 55. The electrospray solution was replaced every second week and capillaries were replaced weekly. The Super SESI solvent reservoir pressure was set to 1.3 bar. The temperature of the ion chamber was set to 90°C and the sampling line temperature was set to 130°C. The exhaust mass flow controller of SUPER SESI was set at 0.7 L/min and nitrogen mass flow through the source was set at 0.4 L/min to ensure a constant fraction of breath entering the ionizer (0.3 L/min).

Quality control

The standardized data collection involved a quality control (i.e. system suitability test, SST) of the analytical platform by using a standard gas-mix prior to measuring the first breath sample of the day. When the stability of the analytical platform was ensured, breath samples were acquired adhering to a previously described protocol (Figure 1). For the quality control 2 ppm of a standard gas-mix (Dalian Special Gases, Dalian, China) containing eight compounds (Table S1) was infused with a flow of 10 L/min using a gas dilution calibrator (Model 2010, Sabio Environmental, Round Rock, USA). The acquired file was then uploaded to a self-developed MATLAB-app (version 2.2) using Hotelling's T^2 (scalar number that summarizes all the score values from PCA) to compare against the historical standard gas-mix signal (Figure S1). For a successful SST the standard gas-mix measurement must a) fall within the 95% CI of the previous values, b) adhere to the Nelson Rules and c) show a signal intensity greater than 1e7 for α-Terpinene. As an additional feature the MATLAB-app provides a comparison of the background ions between the current measurement and previous measurements to indicate potential contamination in the instrument.

QUANTIFICATION AND STATISTICAL ANALYSIS

Untargeted pre-processing of mass spectral data

Pre-processing of exhaled breath data were performed using our patented data processing pipeline (European patent No. 20186274.5 and 21185400.5). Briefly, RAW files were read and converted into MATLAB structure using an in-house C# console app based on an open-source.Net assembly from Thermo Fisher Scientific called RawFileReader (version 5.0.0.38). The time during which CO_2 concentrations was above 3% (measured by the Exhalion) was used to define exhalation time windows. Features from all files were centroid (intensity higher than 1e4), appropriately combined using MATLAB's ksdensity function and correlated with exhalations (i.e. CO_2 trace) to generate a final feature list of size 3198. Then, using the time-trace of each feature, an area under the cure (nAUC) normalized to exhalation time was calculated to depict breath-levels of the respective feature. Later, the final data matrix was transformed to fifth root to reach normally distributed data for further analyses.