Cancer Letters 590 (2024) 216881

Contents lists available at ScienceDirect

Cancer Letters

journal homepage: www.elsevier.com/locate/canlet

Original Articles

Exhaled volatolomics profiling facilitates personalized screening for gastric cancer

Jian Chen [a,1], Yongyan Ji [a,1], Yongqian Liu [a], Zhengnan Cen [a], Yuanwen Chen [b], Yixuan Zhang [b], Xiaowen Li [b,**], Xiang Li [a,*]

[a] Department of Environmental Science & Engineering, Fudan University, Shanghai, 200438, PR China
[b] Department of Gastroenterology, Huadong Hospital Affiliated to Fudan University, Shanghai, 200040, PR China

ARTICLE INFO

Keywords:
Gastric cancer
Exhaled volatolomics
Non-invasive detection
Biomarker

ABSTRACT

Gastric cancer (GC) is one of the most fatal cancers, characterized by non-specific early symptoms and difficulty in detection. However, there are no valid non-invasive screening tools available for GC. Here we establish a non-invasive method that employs exhaled volatolomics and ensemble learning to detect GC. We developed a comprehensive mass spectrometry-based procedure and determined of a wide range of volatolomics from 314 breath samples. The discovery, identification and verification research screened a biomarker panel to distinguish GC from controls. This panel has achieved 0.90 (0.87–0.94, 95%CI) accuracy, with an area under curve (AUC) of 0.92 (0.89–0.94, 95%CI) in discovery cohort and 0.88 (0.83–0.91, 95%CI) accuracy with an AUC of 0.91 (0.87–0.93, 95%CI) in replication cohort, which outperformed traditional serum markers. Single-cell sequencing and gene set enrichment analysis revealed that these exhaled markers originated from aldehyde oxidation and pyruvate metabolism. Our approach advances the design of exhaled analysis for GC detection and holds promise as a non-invasive method to the clinic.

1. Introduction

Gastric cancer (GC) is one of the most fatal cancers around the world, along with a five-year survival rate of less than 30 % [1]. The primary reason is the lack of characteristic symptoms in early-stage GC, leading to many patients to be diagnosed at advanced stage [2]. Early detection and treatment of GC can elevate the five-year survival rate to an impressive 90% [3]. However, current detection methods, such as gastroscopy, pathology, and radiology examinations are not only costly but also carry risks like bleeding and perforation [4]. Therefore, there is an urgent clinical need for a precise and non-invasive tool for the early detection of GC.

Exhaled volatolomics, the characterization of volatile organic compounds (VOCs) in human breath, allows the evaluation of diagnostic and prognostic biomarkers in cancer [5–7]. Gastric cancer can generate numerous specific VOCs that reflect its core pathological features by altering the metabolic mechanism and gastrointestinal microbiota composition of the body. Extensive research has confirmed the significant potential of exhaled VOCs as biomarkers for the non-invasive detection of GC [8–10]. However, further translation of this approach into the clinic has been delayed for two main reasons. Firstly, the scarcity of robustly powered clinical trials, combined with a lack of standardized procedures for the collection, detection, analysis, and processing of exhaled VOCs results in poor stability and reproducibility of breath test results [11]. Secondly, owing to the complex physiological processes and metabolic pathways of exhaled biomarkers, their potential internal origin has long been debated, hindering their clinical application [12,13].

In response to these challenges of exhaled biomarkers in the clinical detection of GC, our study constructed a comprehensive mass spectrometry-based procedure, composed of thermal desorption gas chromatography with triple quadrupole mass spectrometer (TD-GC-MS/MS). Coupled with the ReCIVA breath sampler, this procedure achieved accurate qualitative and quantitative analysis of 82 exhaled VOCs. In addition, we developed an ensemble learning-based framework, taking unique advantage of six machine learning algorithms, which

* Corresponding author.
** Corresponding author.
E-mail addresses: xiaowenli12@fudan.edu.cn (X. Li), lixiang@fudan.edu.cn (X. Li).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.canlet.2024.216881
Received 16 January 2024; Received in revised form 2 April 2024; Accepted 9 April 2024
Available online 12 April 2024

successfully identified a biomarker panel of six exhaled VOCs for non-invasive screening of GC. Furthermore, we performed single-cell sequencing to investigate the generation mechanism of aldehyde and short-chain fatty acid markers in GC patients, which revealed potential internal sources of GC exhaled biomarkers and provided a robust theoretical basis for their application in clinical settings.

2. Materials and methods

2.1. Overview of participants and samples

A total of 157 participants, including 73 gastric cancer patients and 84 matched controls (including 20 gastritis), were ultimately selected for the final analysis. The detailed subject recruitment process was shown in Fig. S1. Subsequently, the 157 subjects underwent post hoc randomization into discovery and replication cohorts in a 2:1 ratio, employing block random assignment. This randomization process was meticulously stratified based on three key criteria: (i) the adjudicated clinical diagnosis, (ii) the interval from hospital admission to breath-testing, and (iii) the clinical diagnostic uncertainty score, facilitated by the R package 'randomizr'. This process resulted in 104 subjects in the discovery cohort and 53 in the replication cohort. All participants were recruited from Huadong Hospital, affiliated with Fudan University in Shanghai, between July 2023 and December 2023. The majority of the participants were from Shanghai or the Yangtze River Delta region. Thus, their dietary preferences leaned towards lighter options with relatively minimal variations in diet patterns. Participants were excluded based on pathological criteria if they: (i) had a history of other malignant tumors before diagnosis or had token neoadjuvant therapy, (ii) suffered from any respiratory diseases, and (iii) were pregnant, breastfeeding, or consumed alcohol daily. Informed consent was obtained from each participant before their participation in the study, which was conducted in strict compliance with the Declaration of Helsinki and received approval from the Ethics Committee of Huadong Hospital (KY 2023K127). Moreover, we added 30 colorectal cancer patients (CRC) to further explore and validate the specificity of our identified exhaled biomarkers. These patient data were derived from other research projects recently focused on the same topic by our research group.

For this study, control volunteers were defined as individuals without significant gastrointestinal disease history. This specifically excluded those diagnosed with gastric ulcers or any form of gastrointestinal cancer. Additionally, individuals with a family history of gastric cancer or who had undergone major gastrointestinal surgery were not included. We decided to include participants with mild or intermittent gastric discomfort but excluded those presenting with severe or persistent symptoms of gastritis.

2.2. Breath sampling methodology

Once obtaining informed consent from all patients, we strictly followed a standardized sampling procedure using an ReCIVA sampler comprised of breath biopsy cartridges and a portable air supply for exhaled sample collection. To minimize the interference of dietary related confounding factors, we performed sample collection between 7:00 and 8:00 a.m. after an overnight fast. Patients were also asked to rest in the same area for at least 20 min and to rinse their mouth three times with clean water before sampling. For each participant, we collected two parallel samples of 2 L alveolar breath gas with corresponding ambient samples. Target VOCs were collected in two duplicate multi-layer thermal desorption (TD) tubes containing Carbograph 5 TD and Tenax/TA (Markes biomonitoring tubes, Markes International Ltd, UK).

2.3. Instrumental analysis

This study used the TD-GC-MS/MS system to analyze the exhaled VOCs of participants. The performance indicators of TD-GC-MS/MS was shown in Table S1. Detailed processes and parameters can be found in the supplementary materials.

2.4. VOCs identification and quantitation

The chemical characteristics of each peak were confirmed by reference to the National Institute of Standards and Technology (NIST) mass spectral library (version 2.3). After confirming the retention time and mass spectrum of the target compounds in SCAN mode, quantitative analysis was performed in Selected Ion Monitoring (SIM) and Multiple Reaction Monitoring (MRM) modes. The Agilent Mass Hunter quantitative analysis software and the Agile2 integrator were used to automatically integrate compound peaks, with manual adjustments made as necessary. A combination of external standard curves and internal standard normalization was used to quantify 82 VOCs.

2.5. Construction of baseline models

We employed six different machine learning algorithms to generate risk predictions for malignant tumors and construct baseline models [14]. These machine learning algorithms included: Rule-based C5.0 (C5), Naive Bayes (NB), Multivariate Adaptive Regression Splines (MARS), Polynomial Support Vector Machine (SVM), Extreme Gradient Boosting Trees (XGB) and Random Forest (RF). To construct each baseline model, we randomly split the dataset into a training set (75 %) and a test set (25 %). We trained each model through 25 bootstrap resamples and used Latin hypercube sampling with a grid size of 50 to tune the hyperparameters of each model. The AUC values and accuracy of the ROC curve for each hyperparameter combination were calculated in each bootstrap resample. Specifically, each baseline model obtained optimal hyperparameter combinations derived from 50 training selections and averaged results from 25 bootstrap iterations [15].

2.6. Establishment of a meta model using LASSO

We constructed the meta-model using LASSO regression based on a stacking strategy coupled with a 6-dimensional feature vector. This stacking approach facilitates the automatic exploration of different baseline models distinct from models built using direct integration strategies such as majority voting and average scoring [14,16–18]. By intelligently integrating their respective strengths without manual intervention, the final meta-model offers improved and more stable performance. Additionally, LASSO regression allows for a statistically sound feature importance analysis, quantifying the impact of each baseline model on the final meta-model's performance. This advantage allows us to more accurately delineate and comprehend the contribution of each VOC feature and individual baseline model to the enhancement of the final stacking model performance.

2.7. Feature selection

To identify exhaled biomarkers with significant information for GC classification, we calculated the importance weight of each VOC feature using the stacking model. After obtaining the weight of each feature, we proceeded with a feature selection process based on the greedy algorithm, analyzing the VOC features one by one according to their descending weights [19]. The greedy feature selection began with an empty set, subsequently adding features based on their potential to enhance classification accuracy. When considering the nth feature, our greedy selection approach firstly put it into the set of features that were previously selected. This was followed by a thorough evaluation, utilizing ten-fold cross-validation repeated 10 times, to gauge the stacking

model's average performance. If this evaluation indicated a performance improvement over the set of previously chosen features, the nth feature was retained in the final selection. Otherwise, it was excluded.

2.8. Single-cell quality control and data processing

We obtained the single-cell samples from GSE167297 [20]. The detailed single-cell data processing process and subsequent analysis can be found in the supplementary materials.

2.9. Statistical analysis

Statistical analysis was conducted using SPSS version 26 (IBM, Armonk, New York, USA) and RStudio (version 4.2.3, RStudio Inc., Boston, MA, USA). In SPSS, univariate non-parametric tests (Wilcoxon signed-rank) were used to evaluate the differences in exhaled VOC levels between gastric cancer patients and the control group, with a P value $<$ 0.05 indicating statistical significance. We used Spearman correlation analysis to investigate the association between respiratory markers and traditional serum markers. In RStudio, Principal component analysis (PCA) was employed for dimensionality reduction and clustering of VOC data.

3. Results

3.1. Overview of the exhaled volatolomics landscape of GC

To obtain the exhaled volatolomics landscape of GC, we collected 314 breath samples from 73 GC patients and 84 matched controls (served as NC), chosen based on pathological criteria (see Materials and methods). The clinicopathological features, including age, sex, gender, TNM classification, and tumor diameter were summarized in Fig. 1A and Table S2. Fig. 1A illustrates the sample distribution according to the TNM classification, which was further divided into early-stage patients (n = 26), advanced-stage patients (n = 40), and patients without staging information (n = 7).

Fig. S2 shows a schematic of the experimental design. We used the ReCIVA sampler to collect the breath samples with ambient samples. After sample collection, a comprehensive MS-based absolute quantification strategy composed of TD-GC-MS/MS was performed to characterize exhaled VOCs from these samples. Analysis of exhaled VOC data revealed 86 unique chromatographic feature peaks, among which we achieved accurate qualitative and quantitative analysis of 82 VOCs, including hydrocarbons (31.33 %), acids (25.81 %), aromatics (16.21 %), aldehydes (10.03 %), ketones (4.00 %), alcohols (3.48 %), esters (1.90 %), sulfur compounds (1.80 %) and others (5.44 %) (Fig. S3 and Table S3). Compared to previous breath-based studies using the MS-based method, this study identified a greater variety of VOC types to achieve a comprehensive exhaled volatolomics landscape of GC [21–28]. Principal component analysis (PCA) of VOC data demonstrated a basic separation between the GC and NC at the exhaled volatolomics level (Fig. 1B). Furthermore, we observed obviously higher butyric acid and valeric acid than in NC. In contrast, hexanoic acid, 2-methyl-hexane, decane and undecane were found lower level than in NC (Fig. 1C and D). These cumulative findings enhanced our understanding of the molecular mechanisms of GC through the exhaled volatolomics landscape.

3.2. Ensemble learning-based selection of biomarkers

We developed a novel ensemble learning framework named Prioritization of Optimal Biomarker Panel for Gastric Cancer (POB-GC) to extract specific VOC features from discovery cohort to identify potential biomarkers. This framework contained three key steps: (1) construction of baseline models; (2) integration of baseline models by stacking strategy; (3) application of greedy algorithm for VOC feature selection (Fig. 2A). We utilized six machine learning algorithms on the volatolomics dataset to generate 150 baseline models (6*25) by applying a 0.5 threshold with 25 bootstrap resampling (see Materials and methods). These formulated models demonstrated an average AUC exceeding 0.74 (Fig. S4). Despite exhibiting decent performance, these baseline models did not outperform those in previous studies that employed classical machine learning algorithms, evidenced by relatively lower performance metrics [25,26]. Such discrepancy was attributable to these models' optimal performance with extensive datasets and their struggle to extract valuable information from limited datasets [29]. To address this, we integrated these baseline models into a hybrid ensemble learning model through stacking strategy implemented by the LASSO regression algorithm. Fig. S5 displays the respective proportions of six machine learning algorithms comprising the stacking model, with the xgboost algorithm having the highest proportion of 55.56 %.

To precisely assess the efficacy of this model, we conducted three sets of experiments, corresponding to three distinct ensemble learning models derived from amalgamating all 150 baseline models: the stacking model, the average scoring model, and the majority voting model [30–32]. Fig. 2B and C clearly demonstrate that the stacking model outperformed the baseline models in all metrics (including SP, SN, ACC, MCC, F-value, and AUC, confirming that the use of stacking strategy can enhance the predictive performance of individual models, consistent with numerous previous studies [33,34]. Furthermore, the stacking model proved to be a superior classifier than the average scoring model and majority voting model, achieving the highest metrics combination for SP (0.92), SN (0.96), ACC (0.94), MCC (0.89), F-value (0.94), and AUC (0.95) (Fig. 2D and E). This reaffirmed that the stacking model could effectively take advantage of individual baseline models for more stable and accurate GC prediction.

We then used the entire exhaled volatolomics dataset as input for the stacking model and screened the most significant biomarkers utilizing the feature weights provided by this model. Initially, we performed 10 times of ten-fold cross-validation to obtain the mean weights of each VOC feature, which were then ranked in descending order. The greedy algorithm was sequentially used to evaluate the top-ranking VOC features. For each feature, if its combination with the previously selected features realized enhanced performance, this feature would be involved in the selected VOC feature set. After feature selection, it was observed that when the number of VOC features increased to six, the accuracy of this model reached 0.90 without significant further improvement (Fig. S6). Therefore, we confirmed the top six VOCs with feature weight (including propionic acid, cyclohexane, heptanal, butyric acid, valeric acid, and undecane) as the specific exhaled biomarker panel for detection of GC. Multiple logistic regression revealed these six VOCs biomarkers were independent predictor for GC (P value < 0.01, $\beta \neq 0$, 95 % CI) (Table S4). Importantly, the combination of six biomarkers together accounted for an enhanced AUC of 0.92 by multivariate ROC curve analysis, compared to the poor diagnostic performance by single one of these biomarkers (AUC $<$0.8) (Fig. S7).

3.3. Superior performance verified in the replication cohort

To assess the potential for clinical application, we evaluated the performance of the exhaled biomarker panel for GC detection in the replication cohort. We recruited 10 cancer-free participants from medical examination as NC and 43 patients including 20 gastritis patients (served as NC) and 23 gastric cancer patients (served as GC). We examined the expression of exhaled biomarkers in this cohort. The Wilcoxon signed-rank test revealed increased propanoic acid expression in GC and elevated heptane in NC (P value < 0.05), indicating that exhaled biomarkers could differentiate GC from NC in replication cohort (Fig. 3A). This panel achieved an AUC of 0.91 with 0.88 accuracy, 0.87 specificity and 0.90 sensitivity in detecting GC from healthy controls (Fig. 3B), accurately identifying 89 % (8 of 9) early GC (stage I and II) with 0.92 AUC and 86 % (12 of 14) advanced GC (stage III and IV) with

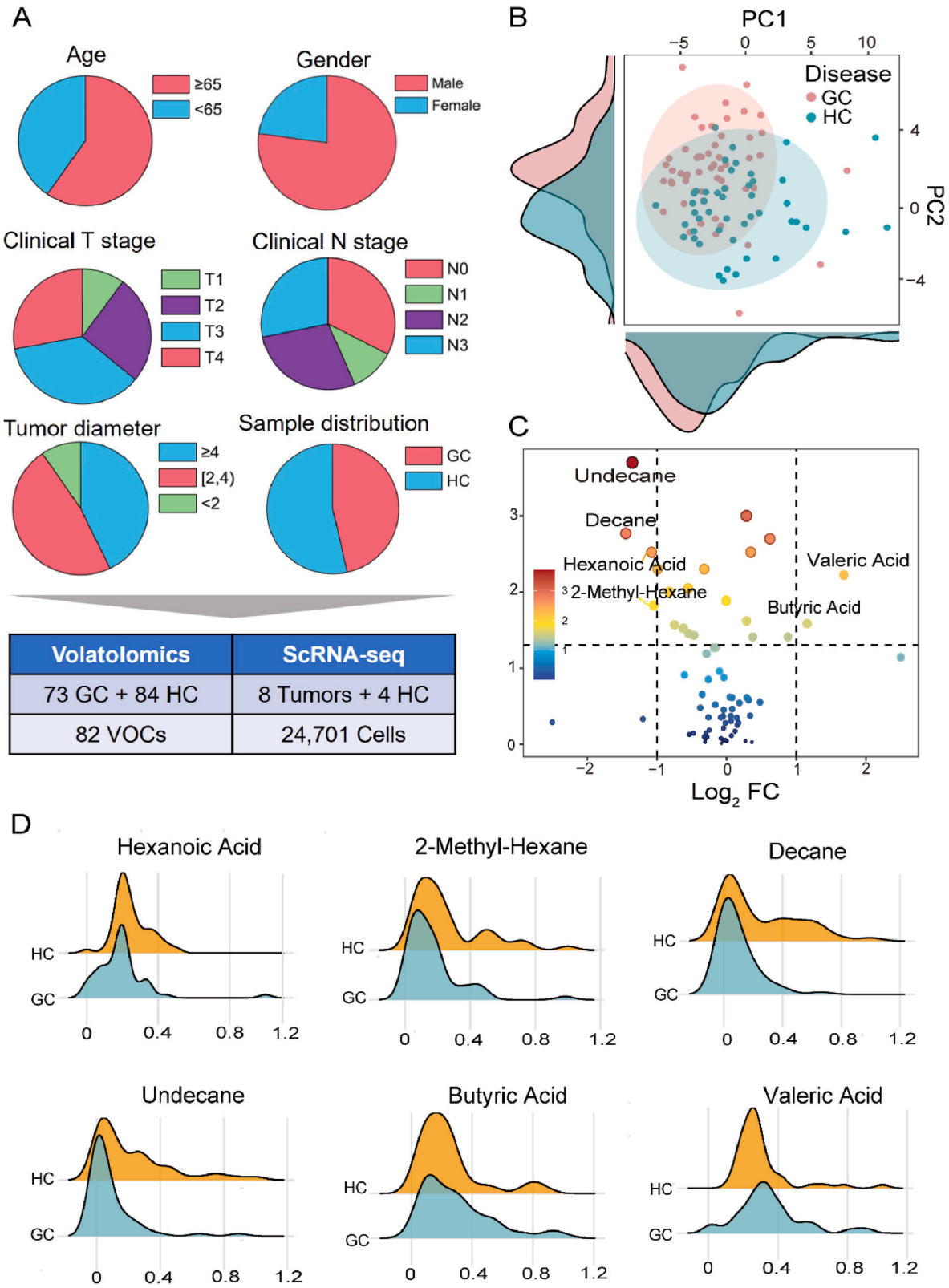

Fig. 1. Exhaled Volatolomics landscape of GC. **A** Top panel, pie charts of clinical indicators. Bottom panel, sample numbers and multi-omics datasets. **B** The PCA analysis of the 82 VOCs between GC and NC. **C** Volcano plots displaying the differentially expressed VOCs in GC and NC after applying a two-fold change in expression with $P < 0.05$ (Wilcoxon rank-sum test). **D** The general density curves visualizing concentration distribution of six differentially expressed VOCs. The x-axis corresponded to normalized concentration of individual VOC and the y-axis corresponded to frequency density.

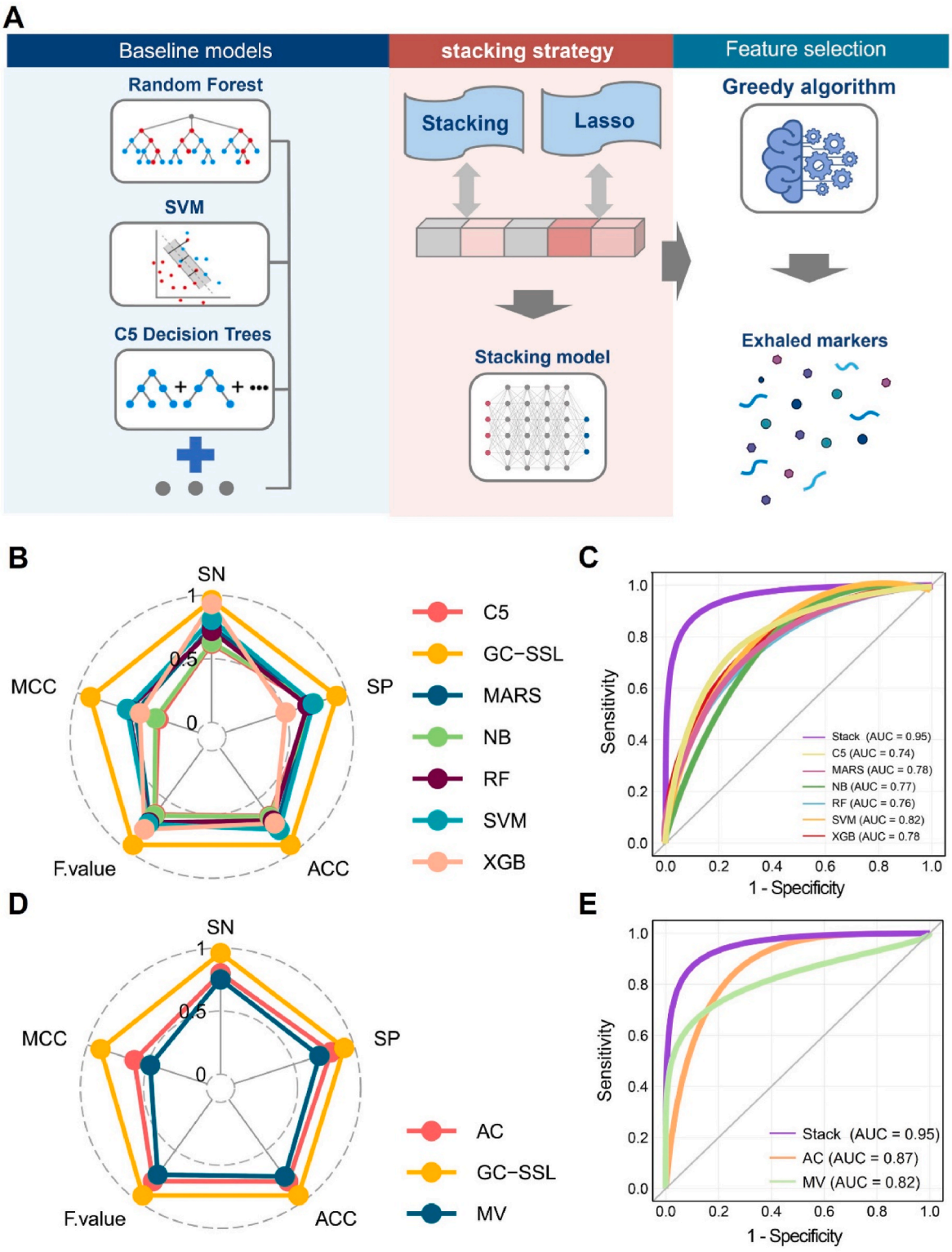

Fig. 2. Identification of potential exhaled biomarkers of GC using POB-GC. **A** The workflow of POB-GC, including (1) construction of baseline models; (2) integration of baseline models by the stacking strategy; (3) application of the greedy algorithm for VOC feature selection. **B** Radar plot showing the SN, SP, ACC, F-value and MCC of the stacking model and six baseline models. **C** ROC for GC detection based on the stacking model and six baseline models. **D** Radar plot showing the SN, SP, ACC, F-value and MCC of the stacking model, average scoring model (AC) and majority voting model (MV). **E** ROC for GC detection based on the stacking model, AC and MV.

0.89 AUC (Fig. S8). It is worth noting that exhaled markers could distinguish GC from gastritis patients, achieving 0.82 specificity, 0.85 sensitivity and 0.86 AUC. Additionally, to further explore the diagnostic specificity of exhaled markers for GC, this study incorporated an additional cohort of 30 cases with colorectal cancer (CRC). The result demonstrated that exhaled biomarkers are capable of distinguishing GC from CRC, with 0.79 specificity, 0.84 sensitivity and 0.83 AUC (Fig. S9).

We then compared the diagnostic performance of the exhaled

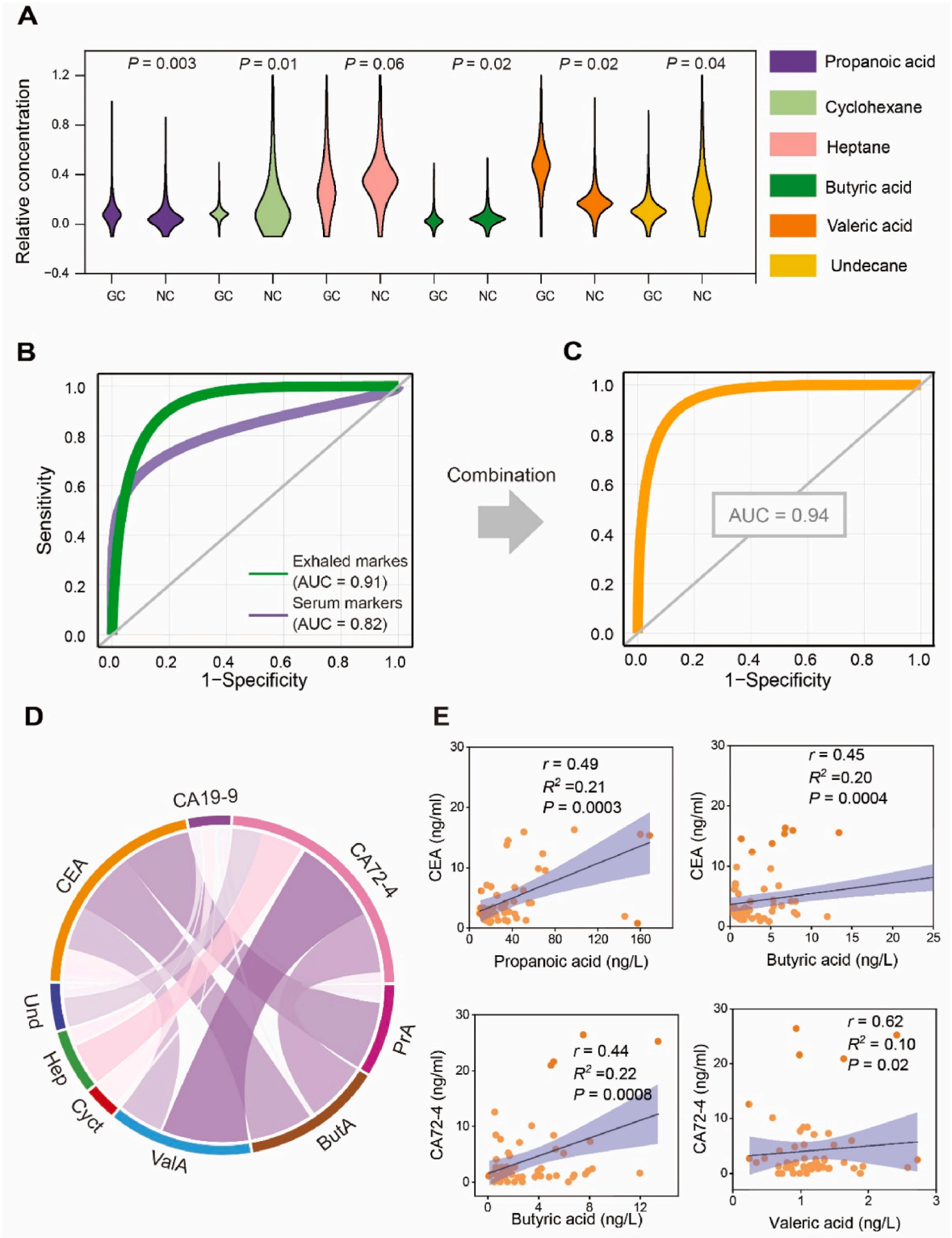

Fig. 3. The performance verification of exhaled markers in replication cohort. **A** Violin plot of exhaled marker expression in replication cohort. **B** ROC for GC detection based on exhaled markers and serum markers. **C** ROC for GC detection based on the combination of two types of markers. **D** Circle plot showing the correlation between exhaled markers and serum markers. Purple representing a positive correlation and pink representing a negative correlation. PrA propionic acid, ButA butyric acid, ValA valeric acid, Cyct cyclohexane, Hep heptanal, Und undecane. **E** The linear fit analysis showing the association of three exhaled markers and two serum markers. *R* correlation coefficient, R^2 coefficient of determination.

biomarkers with the typical serum biomarkers including CEA, CA19-9 and CA72-4. Fig. 3B shows that the AUC of serum biomarkers was 0.82, and the AUC of exhaled biomarkers combined with serum biomarkers could be further improved to 0.94, which indicated this combination improved the performance for detection of GC (Fig. 3C). Spearman correlation analysis confirmed that exhaled biomarkers (propionic acid, butyric acid and valeric acid) correlated with serum biomarkers (CEA and CA72-4) ($r > 0.3$, P value < 0.05) (Fig. 3D and E).

These results demonstrated the ideal performance of the exhaled biomarker panel in distinguishing GC from NC. Therefore, the exhaled biomarker panel could be a promising and non-invasive approach for GC detection in clinics, and the combination of exhaled and serum biomarkers could further improve the clinical screening of GC.

3.4. Unraveling the mechanisms behind biomarker generation

To explore the generation mechanism of exhaled biomarkers, we investigated their potential sources in GC through transcriptomic profiling and exhaled volatolomics analysis. Single-cell RNA sequencing (scRNA-seq) analysis was first performed on a public dataset of GC (see Materials and Methods). We gathered eight GC tissue samples and four normal gastric tissue from four GC patients. From each patient, three samples were collected: one from the superficial layer of the tumor site, the other from the deep layer of the tumor site, and the third from a normal stomach tissue. After quality filtering, 21,521 cells were detected, with a median of 1321 genes per cell, of which 18,760 and 3180

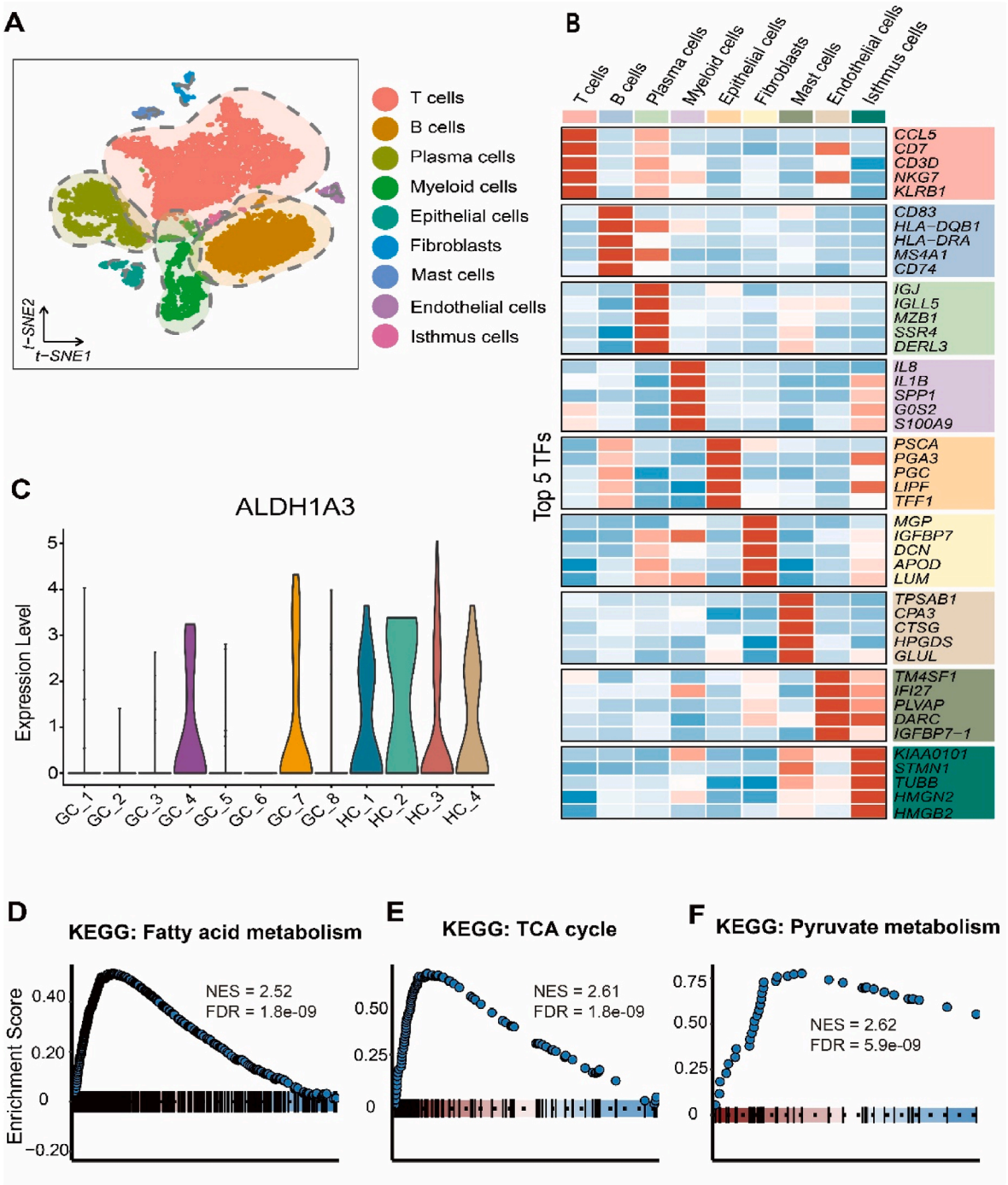

Fig. 4. Single-Cell RNA-Seq analyses of GC and normal gastric tissues. **A** Cell populations identified in human gastric tissues. The t-SNE projection of 24,701 single cells from GC samples(n = 8) and control gastric samples (n = 4). Nine major cell clusters identified are labelled. Each dot corresponds to a single cell and is colored according to its cell type. **B** Heat map displaying DEGs for each cluster. The signal indicates the average expression for each cluster. Representative genes are displayed on the right. **C** Violin plots showing the expression levels of ALDH1A3 in different samples **D-F** Cluster-specific enrichment of malignant cells evaluated by GSEA. NES normalized enrichment score, FDR false discovery rate.

cells were collected from GC tissues and normal gastric tissues (served as NC), respectively. To generate a comprehensive view, we categorized cells into 9 major lineages, including epithelial cells, endothelial cells, mast cells, plasma cells, myeloid cells, fibroblasts, isthmus cells, B cells and T cells (Fig. 4A) based on well-established marker genes (Fig. 4B and Fig. S10). As illustrated in Fig. S11, GC samples exhibited increased proportions of T cells, myeloid cells and isthmus cells, whereas decreased proportions of epithelial cells and endothelial cells, compared to NC samples. We then used large-scale copy number variations (CNVs) to distinguish GC cells from normal cells [35,36]. Epithelial cells exhibited significantly higher CNVs than other cells, indicating epithelial cells were highly malignant (Figs. S12 and S13). Thus, epithelial cells were served as tumor cells in GC, and as normal cells in NC.

We directly compared malignant cells and normal epithelial cells across the 117 metabolism pathways. The results revealed significant disorder not only in several typical tumor metabolic pathways including oxidative phosphorylation, cell cycle, chemical carcinogenesis-reactive oxygen species (ROS), and carbon metabolism but also in some VOC-related metabolism pathways comprising fatty acid metabolism, pyruvate metabolism, and the tricarboxylic acid cycle. Moreover, we observed ALDH1A3 gene deletion in GC samples, which had been proved to lead to an increase in endogenous aldehyde in the exhaled breath of cancer patients (Fig. 4C) [37]. Therefore, we infer that the absence of ALDH1A3 results in the formation of heptanal in GC patients. We then characterized these VOC-related metabolic disorder by comparing GC cells to NC ones, and ranking these pathways by normalized enrichment scores (NES) (Fig. 4D–F). These analyses demonstrated that pyruvate metabolism was the most remarkably variable VOC–related metabolic pathway in GC cells (Fig. 4F). Pyruvate metabolism was up-regulated in GC cells compared with normal cells, accounting for the increase of butyric acid and valeric acid in our volatolomics data. In addition, we infer that cyclohexane and undecane in exhaled biomarkers generated during the lipid peroxidation effect of reactive oxygen species on cell membrane polyunsaturated fatty acids [38–40]. Thus, these scRNA-seq and exhaled volatolomics results demonstrated exhaled biomarkers originated from aldehyde oxidation and pyruvate metabolism. In conclusion, we have proved the characteristic of disturbed VOC-related metabolism of GC and constructed an exhaled biomarker panel based on volatolomics coupled with ensemble learning, and this panel can effectively detect GC.

4. Discussion

In this pragmatic, non-invasive study, we constructed a comprehensive MS-based procedure, composed of a SIM and MRM array, coupled with ReCIVA breath sampler to investigate the potential for identifying VOCs in patient breath. The improved breath sampler was composed of breath biopsy cartridge and portable air supply, which together provide minimal contamination from external VOCs to reduce background noise. Using this sampler, we collected 2 L alveolar breath gas for each participant along with ambient samples between 7:00 and 8:00 a.m., after an overnight fast. This standardized breath sampling procedure minimized the interference of confounding factors (including diet, ambient gas, sampling time, etc.), greatly improving the stability and reliability of the collected results. The TD-GC-MS/MS system was used to enable accurate qualitative and quantitative analysis of 82 exhaled VOCs with high resolution and sensitivity. The types of VOCs identified by this system exceed the results of the current breath-based GC studies using MS-based procedure. These advantages provide a reliable data analysis for the subsequent screening program of exhaled biomarkers.

Traditionally, biomarker identification in metabolomics or exhaled volatolomics mainly relied on statistical methods (such as variance analysis and partial least squares discriminant analysis) and simple machine learning models (such as random forests and decision trees) [8,41,42]. In this study, we innovatively employed the stacking strategy-based ensemble learning model, which exhibited excellent performance in detecting GC patients. This success can be attributed to the stacking strategy which integrates the unique advantages of six machine learning algorithms, comprehensively capturing subtle features in the exhaled volatolomics dataset [17]. On the basis of the ensemble learning model, an exhaled biomarker panel was established using the greedy algorithm. Considered as an emerging cancer screening method, the performance of exhaled biomarkers is more precise and non-invasive than those from serum metabolomics. The successful identification of this panel from exhaled volatolomics coupled with stacking model and greedy algorithm provides a valuable insight into how to promote exhaled VOCs as biomarkers for the early detection of GC.

Current biomarker-based screening approaches available for GC primarily depend on traditional serum biomarkers such as CEA, CA19-9, and CA72-4 [43,44]. Apart from these, no other biomarkers have been widely accepted and applied in the clinic. However, these markers exhibit certain limitations, including an increase of false positives in patients with benign gastric disease, owing to lack of specificity for GC. Our replication cohort study demonstrated that exhaled biomarkers can accurately detect GC from NC. Moreover, the AUC of exhaled biomarkers could be further improved to an astonishing 0.94 when coupled with serum biomarkers, suggesting that exhaled biomarkers complement the diagnostic shortcomings of serum biomarkers. Our exhaled biomarkers could not only serve as an independent signature to detect GC but could also be combined with serum biomarkers to provide clinicians with a more comprehensive screening tool, assisting them in arranging more appropriate subsequent diagnostic and treatment procedures.

Among these selected biomarkers, there were three types of short-chain fatty acids (propionic acid, butyric acid and valeric acid), two types of alkanes (cyclohexane and undecane) and one type of aldehyde (heptanal). ScRNA-seq analyses revealed that pyruvate metabolism was the most significantly up-regulated VOC-related metabolic pathway in GC cells, accounting for the increase of butyric acid and valeric acid in GC patients. Moreover, we observed ALDH1A3 gene deletion in GC samples, which resulted in heptanal increase in GC patients by restraining of aldehyde oxidation [37]. In addition, we infer that cyclohexane and undecane in exhaled biomarkers are generated from the lipid peroxidation effect of reactive oxygen species on cell membrane polyunsaturated fatty acids [38–40]. We have investigated potential sources of exhaled biomarkers in GC by exploring their generation mechanism, providing a theoretical basis for their clinical application.

Certain limitations of this study should be recognized. Owing to the current limited sample size, we were unable to pool and label patients with GC at different stages for training the stacking model. Thus, the signatures identified by the stacking model were regarded as common characteristics of early and advanced GC, which could be further revised as large-scale sampling is in progress. The exhaled biomarker panel was constructed and validated in discovery cohort and replication cohort, both recruited from Huadong Hospital affiliated to Fudan University in Shanghai. The generality of this panel for GC screening in other populations requires further investigation. Given the uncertain relationships between GC and diabetes or obesity, the performance of this panel could be influenced by metabolic-related confounders [45]. Before being officially recognized as an early diagnosis approach for GC, this biomarker panel must undergo extensive analysis and clinical validation in multi-ethnic, multi-center, and large-scale cohorts with strict enrollment criteria.

In conclusion, we have established an exhaled biomarker panel through ensemble learning coupled with a greedy algorithm to improve the disease detection process guided by exhaled volatolomics. The advantages of this panel underscore its potential application in biomarker-aided detection of GC. This precise and non-invasive approach offers a novel prospect for disease screening through exhaled volatolomics. We believe that the appropriate clinical application of exhaled biomarkers

could be beneficial to GC patients for accurate detection, resulting in more effective treatment and prognosis.

CRediT authorship contribution statement

Jian Chen: Writing – original draft, Visualization, Conceptualization. **Yongyan Ji:** Writing – review & editing, Writing – original draft, Methodology. **Yongqian Liu:** Writing – review & editing, Methodology. **Zhengnan Cen:** Investigation. **Yuanwen Chen:** Writing – review & editing, Supervision. **Yixuan Zhang:** Writing – review & editing, Investigation. **Xiaowen Li:** Writing – review & editing, Investigation. **Xiang Li:** Writing – review & editing, Writing – original draft, Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 22276038 and 42061134006) and Agilent Technologies Inc. (No. 4956).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.canlet.2024.216881.

References

[1] Z. Wang, Y. Chen, X. Li, Y. Zhang, X. Zhao, H. Zhou, X. Lu, L. Zhao, Q. Yuan, Y. Shi, J. Zhao, Z. Dong, Y. Jiang, K. Liu, Tegaserod Maleate suppresses the growth of gastric cancer in vivo and in vitro by targeting MEK1/2, Cancers 14 (2022) 3592.
[2] P. Sharma, Gastro-oesophageal reflux disease: symptoms, erosions, and Barrett's–what is the interplay? Gut 54 (2005) 739–740.
[3] X.Y. Fu, X.L. Mao, Y.H. Chen, N.N. You, Y.Q. Song, L.H. Zhang, Y. Cai, X.N. Ye, L. P. Ye, S.W. Li, The feasibility of applying artificial intelligence to gastrointestinal endoscopy to improve the detection rate of early gastric cancer screening, Front. Med. 9 (2022) 886853.
[4] E.C. Smyth, M. Verheij, W. Allum, D. Cunningham, A. Cervantes, D. Arnold, E. G. Committee, Gastric cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up, Ann. Oncol. 27 (2016) 38–49.
[5] F. Djago, J. Lange, P. Poinot, Induced volatolomics of pathologies, Nat. Rev. Chem 5 (2021) 183–196.
[6] Y.Y. Broza, I. Braverman, H. Haick, Breath volatolomics for diagnosing chronic rhinosinusitis, Int. J. Nanomed. 13 (2018) 4661–4670.
[7] M.K. Nakhleh, H. Haick, M. Humbert, S. Cohen-Kaminsky, Volatolomics of breath as an emerging frontier in pulmonary arterial hypertension, Eur. Respir. J. 49 (2017) 1601897.
[8] P. Wang, Q. Huang, S. Meng, T. Mu, Z. Liu, M. He, Q. Li, S. Zhao, S. Wang, M. Qiu, Identification of lung cancer breath biomarkers based on perioperative breathomics testing: a prospective observational study, EClinicalMedicine 47 (2022) 101384.
[9] H. Fu, New developments of gastric cancer biomarker research, Nano Biomed. Eng. 8 (2016) 268–273.
[10] A. Shaffie, A. Soliman, X.A. Fu, M. Nantz, G. Giridharan, V. van Berkel, H. A. Khalifeh, M. Ghazal, A. Elmaghraby, A. El-Baz, A novel technology to integrate imaging and clinical markers for non-invasive diagnosis of lung cancer, Sci. Rep. 11 (2021) 4597.
[11] P. Sukul, P. Trefz, J.K. Schubert, W. Miekisch, Advanced setup for safe breath sampling and patient monitoring under highly infectious conditions in the clinical environment, Sci. Rep. 12 (2022) 17926.
[12] P. Sukul, A. Richter, J.K. Schubert, W. Miekisch, Deficiency and absence of endogenous isoprene in adults, disqualified its putative origin, Heliyon 7 (2021) e05922.
[13] X.G. Li, J. Du, J. Chen, F. Lin, W. Wang, T.T. Wei, J. Xu, Q.B. Lu, Metabolic profile of exhaled breath condensate from the pneumonia patients, Exp. Lung Res. 48 (2022) 149–157.
[14] R. Shaw, A.E. Lokshin, M.C. Miller, G. Messerlian-Lambert, R.G. Moore, Stacking machine learning algorithms for biomarker-based preoperative diagnosis of a pelvic mass, Cancers 14 (2022) 1291.
[15] E.R. DeLong, D.M. DeLong, D.L. Clarke-Pearson, Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach, Biometrics 44 (1988) 837–845.
[16] S. Imangaliyev, J. Schlotterer, F. Meyer, C. Seifert, Diagnosis of inflammatory bowel disease and colorectal cancer through multi-view stacked generalization applied on gut microbiome data, Diagnostics 12 (2022) 2514.
[17] R. Xie, J. Li, J. Wang, W. Dai, A. Leier, T.T. Marquez-Lago, T. Akutsu, T. Lithgow, J. Song, Y. Zhang, DeepVF: a deep learning-based hybrid framework for identifying virulence factors using the stacking strategy, Briefings Bioinf. 22 (2021) 1–15.
[18] M. Azadpour, C.M. McKay, R.L. Smith, Estimating confidence intervals for information transfer analysis of confusion matrices, J. Acoust. Soc. Am. 135 (2014) 140–146.
[19] G.X. Wang, H.T. Yao, Y. Gong, Z.P. Lu, R.F. Pang, Y. Li, Y.Y. Yuan, H.J. Song, J. Liu, Y. Jin, Y.S. Ma, Y.M. Yang, H.G. Nie, G.Z. Zhang, Z. Meng, Z. Zhou, X.Y. Zhao, M. T. Qiu, Z.C. Zhao, K.R. Jiang, Q. Zeng, L.M. Guo, Y.X. Yin, Metabolic detection and systems analyses of pancreatic ductal adenocarcinoma through machine learning, lipidomics, and multi-omics, Sci. Adv. 7 (2021) eabh2724.
[20] H.Y. Jeong, I.H. Ham, S.H. Lee, D. Ryu, S.Y. Son, S.U. Han, T.M. Kim, H. Hur, Spatially distinct reprogramming of the tumor microenvironment based on tumor invasion in diffuse-type gastric cancers, Clin. Cancer Res. 27 (2021) 6529–6542.
[21] M.E. Adam, M. Fehervari, P.R. Boshier, S.T. Chin, G.P. Lin, A. Romano, S. Kumar, G.B. Hanna, Mass-spectrometry analysis of mixed-breath, isolated-bronchial-breath, and gastric-endoluminal-air volatile fatty acids in esophagogastric cancer, Anal. Chem. 91 (2019) 3740–3746.
[22] S. Kumar, J. Huang, N. Abbassi-Ghadi, H.A. Mackenzie, K.A. Veselkov, J.M. Hoare, L.B. Lovat, P. Spanel, D. Smith, G.B. Hanna, Mass spectrometric analysis of exhaled breath for the identification of volatile organic compound biomarkers in esophageal and gastric adenocarcinoma, Ann. Surg. 262 (2015) 981–990.
[23] H. Tong, Y. Wang, Y. Li, S. Liu, C. Chi, D. Liu, L. Guo, E. Li, C. Wang, Volatile organic metabolites identify patients with gastric carcinoma, gastric ulcer, or gastritis and control patients, Cancer Cell Int. 17 (2017) 108.
[24] Y.J. Jung, H.S. Seo, J.H. Kim, K.Y. Song, C.H. Park, H.H. Lee, Advanced diagnostic technology of volatile organic compounds real time analysis analysis from exhaled breath of gastric cancer patients using proton-transfer-reaction time-of-flight mass spectrometry, Front. Oncol. 11 (2021) 560591.
[25] H. Amal, M. Leja, K. Funka, R. Skapars, A. Sivins, G. Ancans, I. Liepniece-Karele, I. Kikuste, I. Lasina, H. Haick, Detection of precancerous gastric lesions and gastric cancer through exhaled breath, Gut 65 (2016) 400–407.
[26] S.R. Markar, T. Wiggins, S. Antonowicz, S.T. Chin, A. Romano, K. Nikolic, B. Evans, D. Cunningham, M. Mughal, J. Lagergren, G.B. Hanna, Assessment of a noninvasive exhaled breath test for the diagnosis of oesophagogastric cancer, JAMA Oncol. 4 (2018) 970–976.
[27] J. Zhang, Y. Tian, Z. Luo, C. Qian, W. Li, Y. Duan, Breath volatile organic compound analysis: an emerging method for gastric cancer detection, J. Breath Res. 15 (2021) 044002.
[28] Y. Hong, X. Che, H. Su, Z. Mai, Z. Huang, W. Huang, W. Chen, S. Liu, W. Gao, Z. Zhou, G. Tan, X. Li, Exhaled breath analysis using on-line preconcentration mass spectrometry for gastric cancer diagnosis, J. Mass Spectrom. 56 (2020) e4588.
[29] C. Angermueller, T. Parnamaa, L. Parts, O. Stegle, Deep learning for computational biology, Mol. Syst. Biol. 12 (2016) 878.
[30] X.W. Chen, J.C. Jeong, Sequence-based prediction of protein interaction sites with an integrative method, Bioinformatics 25 (2009) 585–591.
[31] J.W. Wang, B.J. Yang, A. Leier, T.T. Marquez-Lago, M. Hayashida, A. Rocker, Y. J. Zhang, T. Akutsu, K.C. Chou, R.A. Strugnell, J.N. Song, T. Lithgow, Bastion6: a bioinformatics approach for accurate prediction of type VI secreted effectors, Bioinformatics 34 (2018) 2546–2555.
[32] Y.J. Zhang, S. Yu, R.P. Xie, J.H. Li, A. Leier, T.T. Marquez-Lago, T. Akutsu, A. I. Smith, Z.Y. Ge, J.W. Wang, T. Lithgow, J.N. Song, PeNGaRoo, a combined gradient boosting and ensemble learning framework for predicting non-classical secreted proteins, Bioinformatics 36 (2020) 704–712.
[33] J.E. Lewis, M.L. Kemp, Integration of machine learning and genome-scale metabolic modeling identifies multi-omics biomarkers for radiation resistance, Nat. Commun. 12 (2021) 2700.
[34] M. Kim, C. Chen, P. Wang, J.J. Mulvey, Y. Yang, C. Wun, M. Antman-Passig, H.-B. Luo, S. Cho, K. Long-Roche, L.V. Ramanathan, A. Jagota, M. Zheng, Y. Wang, D. A. Heller, Detection of ovarian cancer via the spectral fingerprinting of quantum-defect-modified carbon nanotubes in serum by machine learning, Nat. Biomed. Eng. 6 (2022) 267–275.
[35] A.P. Patel, I. Tirosh, J.J. Trombetta, A.K. Shalek, S.M. Gillespie, H. Wakimoto, D. P. Cahill, B.V. Nahed, W.T. Curry, R.L. Martuza, D.N. Louis, O. Rozenblatt-Rosen, M.L. Suva, A. Regev, B.E. Bernstein, Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma, Science 344 (2014) 1396–1401.
[36] I. Tirosh, B. Izar, S.M. Prakadan, M.H. Wadsworth, D. Treacy, J.J. Trombetta, A. Rotem, C. Rodman, C. Lian, G. Murphy, M. Fallahi-Sichani, K. Dutton-Regester, J.R. Lin, O. Cohen, P. Shah, D. Lu, A.S. Genshaft, T.K. Hughes, C.G.K. Ziegler, S. W. Kazer, A. Gaillard, K.E. Kolb, A.C. Villani, C.M. Johannessen, A.Y. Andreev, E. M. Van Allen, M. Bertagnolli, P.K. Sorger, R.J. Sullivan, K.T. Flaherty, D. T. Frederick, J. Jane-Valbuena, C.H. Yoon, O. Rozenblatt-Rosen, A.K. Shalek, A. Regev, L.A. Garraway, Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq, Science 352 (2016) 189–196.
[37] S. Antonowicz, Z. Bodai, T. Wiggins, S.R. Markar, P.R. Boshier, Y.M. Goh, M. E. Adam, H. Lu, H. Kudo, F. Rosini, R. Goldin, D. Moralli, C.M. Green, C.J. Peters, N. Habib, H. Gabra, R.C. Fitzgerald, Z. Takats, G.B. Hanna, Endogenous aldehyde accumulation generates genotoxicity and exhaled biomarkers in esophageal adenocarcinoma, Nat. Commun. 12 (2021) 1454.

[38] I. Horvath, Z. Lazar, N. Gyulai, M. Kollai, G. Losonczy, Exhaled biomarkers in lung cancer, Eur. Respir. J. 34 (2009) 261–275.
[39] N. Ratcliffe, T. Wieczorek, N. Drabinska, O. Gould, A. Osborne, B.L. Costello, A mechanistic study and review of volatile products from peroxidation of unsaturated fatty acids: an aid to understanding the origins of volatile organic compounds from the human body, J. Breath Res. 14 (2020) 034001.
[40] Y.L. Zou, H.X. Li, E.T. Graham, A.A. Deik, J.K. Eaton, W.Y. Wang, G. Sandoval-Gomez, C.B. Clish, J.G. Doench, S.L. Schreiber, Cytochrome P450 oxidoreductase contributes to phospholipid peroxidation in ferroptosis, Nat. Chem. Biol. 16 (2020) 302–309.
[41] A. Romani, G. Marrone, R. Celotto, M. Campo, C. Vita, C. Chiaramonte, A. Carretta, N. Di Daniele, A. Noce, Utility of SIFT-MS to evaluate volatile organic compounds in nephropathic patients' breath, Sci. Rep. 12 (2022) 10413.
[42] Z. Cen, B. Lu, Y. Ji, J. Chen, Y. Liu, J. Jiang, X. Li, X. Li, Virus-induced breath biomarkers: a new perspective to study the metabolic responses of COVID-19 vaccinees, Talanta 260 (2023) 124577.
[43] Y. Zhu, N. Chen, M. Chen, X. Cui, H. Yang, X. Zhu, J. Dai, Y. Gong, D. Gu, X. Huo, H. Huang, C. Tang, Circulating tumor cells: a surrogate to predict the effect of treatment and overall survival in gastric adenocarcinoma, Int. J. Biol. Markers 36 (2021) 28–35.
[44] J. Yang, H. Su, T. Chen, X. Chen, H. Chen, G. Li, J. Yu, Development and validation of nomogram of peritoneal metastasis in gastric cancer based on simplified clinicopathological features and serum tumor markers, BMC Cancer 23 (2023) 64.
[45] B. Dabo, C. Pelucchi, M. Rota, H. Jain, P. Bertuccio, R. Bonzi, D. Palli, M. Ferraroni, Z.F. Zhang, A. Sanchez-Anguiano, Y.T.H. Pham, C.T.D. Tran, A.G. Pham, G.P. Yu, T.C. Nguyen, J. Muscat, S. Tsugane, A. Hidaka, G.S. Hamada, D. Zaridze, D. Maximovitch, M. Kogevinas, N.F. de Larrea, S. Boccia, R. Pastorino, R.C. Kurtz, A. Lagiou, P. Lagiou, J. Vioque, M.C. Camargo, M.P. Curado, N. Lunet, P. Boffetta, E. Negri, C. La Vecchia, H.N. Luu, The association between diabetes and gastric cancer: results from the Stomach Cancer Pooling Project Consortium, Eur. J. Cancer Prev. 31 (2022) 260–269.