

An Improved Lung Cancer Diagnosis Model Based on Vocs and CT Data with Transcriptome Analysis

Qian Wu¹
Jiajing Sheng¹
Yingchang Zou¹
Yanjie Hu²
Kejing Ying¹
Hao Wan^{1*}
Ping Wang^{1*}

¹Biosensor National Special Lab, Key Lab for Biomedical Engineering of Ministry of Education, Department of Biomedical Engineering, Zhejiang University, Hangzhou, China
²Zhejiang Sir Run Run Shaw Hospital, Department of Medicine, Zhejiang University, Hangzhou, China

Abstract

Purpose Lung cancer (LC) is a leading cause of cancer-related morbidity and mortality globally. Exhaled VOCs have been considered as promising biomarkers for LC diagnosis. However, the accuracy of VOCs for LC diagnosis is not high enough due to the interference from benign pulmonary diseases. This study aims to establish an improved lung cancer diagnosis model with high accuracy in distinguishing lung cancer patients from benign pulmonary patients and healthy controls.

Methods: Herein, numbers of exhaled breath samples were analyzed by TD-GCMS, and the power of discrimination of each VOC was evaluate by ROC curve. Optimization was performed by adding related variables and random forests. To explore the biological relationship between selected VOCs and lung cancer, transcriptome data was analyzed by edgeR and DAVIDE.

Results: The VOCs based diagnosis model was optimized through adding variables. To facilitate the sensor measurement, a five variables model with high was established. Based on transcriptome analysis, lung cancer related metabolic pathways were obtained, and some pathways were consistent with biological metabolic processes which generate VOCs *in vivo*.

Conclusions: Our improved model can accurately discriminate LC patients from other patients and health people, which provides a promising approach for lung cancer diagnosis.

Keywords

Lung cancer; VOCs biomarker; CT data; TD-GCMS; Model optimization; Transcriptome analysis

Introduction

Lung cancer (LC) is continuously a leading cause of cancer-related morbidity and mortality worldwide [1]. However, with early LC diagnosis, LC patients can be cured by surgery treatment and subsequent chemotherapy. Nevertheless, a convenient diagnosis method is still in great demand in clinical applications. Hence, how to develop a good early diagnosis approach is the key requirement in clinical [2]. Volatile organic compounds (VOCs) exhaled by humans contain a lot of information about the condition of the body. In the last decades, an increasing number of studies about detecting and analyzing exhaled VOCs have emerged [3,4], and they considered VOCs as promising biomarkers for non-invasive and convenient LC screening [5]. However, many study samples only contain LC patients and healthy people [6]. These studies are different from the fact in hospital that clinical patients include many other pulmonary non-malignant disease (PNMD) patients. Furthermore, this method of detecting VOCs is limited by its poor accuracy and various VOCs results based on different conditions and methods in different studies [7,8]. To increase the reliability of VOCs in lung cancer diagnosis, expanding sample volume and optimizing diagnosis model are two effective approaches. Zou et al. [9] optimized the diagnosis model through the establishment of training cohort and independent validation cohort to exclude interferences of PNMD. However, outstanding improvements are hardly achieved *via* simply optimizing in the level of VOCs. Other related variables should be introduced into the diagnosis model. Besides, a mass of VOCs as biomarkers are difficult to develop a highly precise sensor which can be applied in clinical LC preliminary screening.

On the other hand, due to a lack of biological evidence between VOCs and LC, the approach of VOCs for LC screening is still hard to be validated. Transcriptome analysis which can explain the source of specific VOCs may solve this problem. The genetic and genomic changes existing

Article Information

DOI: 10.31021/jcro.20192119
Type: Research Article
Journal Type: Open Access
Volume: 2 **Issue:** 1
Manuscript ID: JCRO-2-119
Publisher: Boffin Access Limited

Received Date: 21 December 2018
Accepted Date: 21 January 2019
Published Date: 06 February 2019

*Corresponding author:

Hao Wan, Ping Wang
Department of Biomedical Engineering
Yuquan Campus
Zhejiang University
310027 Hangzhou, China
Tel: +86 571 87952832
Fax: +86 571 87952832
E-mail: wh1816@zju.edu.cn (Hao Wan);
cnpwang@zju.edu.cn (Ping Wang).

Citation: Wu Q, Sheng J, Zou Y, Hu Y, Ying K, et al. An Improved Lung Cancer Diagnosis Model Based on Vocs and CT Data with Transcriptome Analysis. J Cancer Res Oncobiol. 2019 Feb;2(1):119

Copyright: © 2019 Wu Q, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 international License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

in cancer patients have been recognized broad [10], and these changes can be reflected by transcriptome analysis conveniently. Recently, an increasing number of researches were studying the LC transcriptome data from The Cancer Genome Atlas (TCGA, <https://gdc-portal.nci.nih.gov/>). Fidler et al. [11] evaluated transcriptome data from TCGA against serum specimens from lung cancer patients and found some protein biomarkers with inferior survival for patients. Since the biological process of cancer is really complicated, only genomic or proteomic perspective is inadequate to explain it. The study of pathway transform in cancer is critical in understanding the disease [12,13]. Lin et al. [14] identified specific pathways like Liver X receptor activation which possibly indicated important differences in cancer cell metabolism. However, to our knowledge, no study on combining VOCs and transcriptome analysis is reported.

In this study, a large amount of exhaled breath samples was analyzed to select a set of VOCs in all samples for LC diagnosis. A LC diagnosis model was established based on VOCs and optimized through adding CT data and other variables using machine learning. At last, transcriptome analyses as well as subsequently pathway analysis were employed to explain the biological metabolisms of the obtained VOCs.

Materials and Methods

The overall flowchart of our research is shown in Figure 1 which exhibits the VOCs analysis procedure as well as establishment, optimization and verification of LC diagnosis model. All the details will be discussed in the following sections.

Collection of study samples

From 1st January 2009 to 31st December 2016, we continually collected and analyzed exhaled breath samples of 197 LC patients and 70 PNMD patients from Sir Run Run Shaw Hospital, Hangzhou, China, and 178 healthy control samples were collected from Department of Biomedical Engineering, Zhejiang University and Sir Run Run Shaw Hospital (Part of the data was reported in our previous study [9]). LC patients were diagnosed on the basis of MRI or CT characteristics [15] and were confirmed by histology or pathology [16]. Informed consent was obtained from every subject. The approval was obtained from the institutional ethics review committee of Sir Run Run Shaw Hospital, Hangzhou, China (No. 20070525 and ChiCTR-DCD-15007106). The volunteers' statistical numbers of gender, age, smoking status and CT data were shown in Table 1. Different classifications like smoking status and CT data are introduced to diagnosis model in optimization process.

Collection, extraction and analysis of exhaled breath

Before breath sample collection, volunteers were asked to stop consuming food, drinks, and smoking for 12 hours and stay in a ventilated room for 30 minutes. In order to collect the same

standard breath sample, every subject breathed normally, through a disposable mouthpiece, and room air was also collected as background at the same time. The VOCs in the collected gas samples were concentrated by Tenax TA sorption tubes (Sigma-Aldrich, St. Louis, MO, USA) for subsequent analysis immediately without any storage.

Thermal desorption device (TurboMatrix 300TD, PerkinElmer, USA) was used to release the VOCs from Tenax tubes. With the aid of carrier gas, the released VOCs were transferred to Gas Chromatography and Mass Spectrometry (GCMS) for VOCs separation and qualitative detection. GC procedures: initially, the column was heated to 40°C and held for 1 min. After that, the column increased to 250°C with 5°C/min and held for 2 min. MS condition: the temperatures of the interface and ion source in the MS were 250°C and 200°C, respectively. Scanning charge ranged from 45 to 500 mass in scan mode. The solvent cut-off time was 0.4 min.

Chromatographic peaks with slope greater than 500/min and peak area higher than 3000 were selected from GCMS chromatogram. According to the mass spectrometry library (NIST 05 and NIST 05s), [4] the corresponding VOCs with similarity higher than 90% were selected for further analysis. Besides, the retention time, Chemical Abstracts Service (CAS) number and peak area of substances were extracted from the raw data. To eliminate the interference of air background, the peak area of substances subtracted the background air response. The substances which existed in less than 50% samples were discarded to guarantee valid VOCs for analysis.

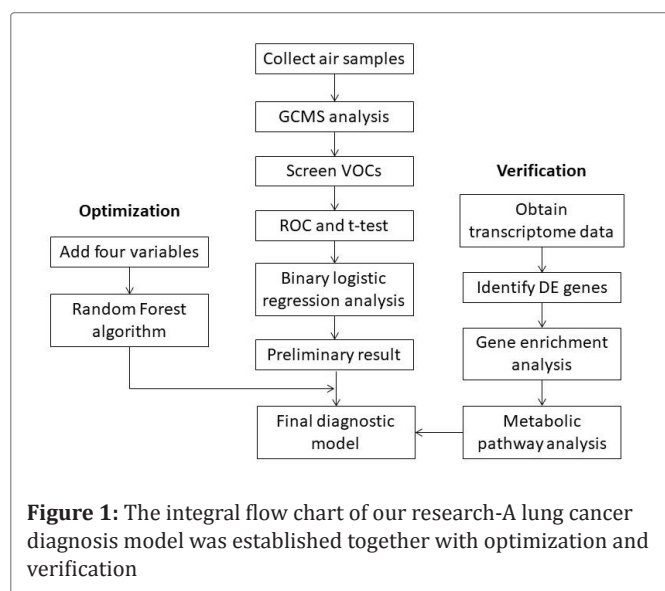
Algorithms for VOCs selection

The receiver operating characteristics (ROC) curve is defined as a plot of test sensitivity as they coordinate versus its 1-specificity as the x coordinate [17]. ROC curve is an effective method of evaluating the performance of qualitative diagnosis, namely a binary result which is either positive or negative for the disease diagnosis. We evaluated the discriminating power of each VOC according to area under the ROC curve (AUC) which can take on any value between 0 and 1, and a VOC with an AUC value of 1 indicates that the VOC is extremely accurate for disease diagnosis. Furthermore, t-test was employed to determine if a VOC has statistically significant difference between LC patients and other people. In dichotomous problem, logistic regression was used to assess the likelihood of falling into one of the outcome categories based on a set of predictors [18]. Herein, binary logistic regression analysis was applied to establish a diagnosis model for LC diagnosis based on VOCs results from previous analysis. Random forests (RF), as a powerful machine learning model with high accuracy in statistical classifier was used to optimize the diagnosis model [19]. Moreover, it is worth stressing that RF enables determining variable importance by mean decrease accuracy (MDA), which can be used to decrease variables for convenient sensing in clinical applications without sacrificing the accuracy of the model.

Bioinformatics analysis

In order to explore the biological source of VOCs in LC patients, we obtained RNA sequencing data, including LC patients and healthy control from TCGA. To get higher accurate differentially expressed genes (DEGs) results, more healthy control cases are required for comparison. Hence, we downloaded RNA-Seq gene expression data for which all alive cases were available in TCGA database.

EdgeR package from R programming language was used for transcriptome data analysis to identify DEGs between cancer samples and healthy samples [20]. This classical method is based on the negative binomial distributions, including empirical Bayes estimation, exact tests, generalized linear models and quasi-likelihood tests. We used false discovery rate (FDR) to define DEGs' threshold by the Benjamini and Hochberg (BH) method. Genes with FDR less than 0.001 and mean gene expression fold-change larger than 4 were identified as the DEGs. For gene enrichment analysis, DAVID 6.8 was used for pathway analysis of gene sets [21]. The pathway analysis provides a comprehensive functional annotation to understand the biological significance behind large list of genes.



Samples (N)	Gender (N)	Age (mean $\pm$ std)	Smoking status	CT data N (%)
LC (197)	Male: 114 Female: 83	63.2 $\pm$ 9.26 59.5 $\pm$ 11.0	Nonsmoker: 85 (43.1%) Ex-smoker: 58 (29.5%) Smoker: 54 (27.4%)	Normal: 59.5 $\pm$ 11.0 Nodule: 26 (13.2%) Tumor: 169 (86.8%)
PNMD (70)	Male: 46 Female: 24	55.6 $\pm$ 12.5 57.5 $\pm$ 12.1	Nonsmoker: 33 (47.2%) Ex-smoker: 16 (22.8%) Smoker: 21 (30.0%)	Normal: 2 (2.9%) Nodule: 47 (67.1%) Tumor: 21 (30.0%)
Healthy control (178)	Male: 142 Female: 36	49.7 $\pm$ 8.5 51.1 $\pm$ 10.2	Nonsmoker: 55 (30.9%) Ex-smoker: 16 (9.0%) Smoker: 107 (60.1%)	Normal: 106 (60.1%)
Total (445)	Male: 302 Female: 143	55.7 $\pm$ 11.3 57.0 $\pm$ 11.4		

Table 1: Statistical information regarding all volunteers

Results and Discussion

VOCs selection

According to the aforementioned screening method, VOC profiles of every participant were obtained, and the statistical numbers are shown in Table 2. On average, no significant difference of VOC numbers was observed from LC, PNMD and healthy control. In order to discriminate LC patients, we divided all samples into LC group and non-LC group (including PNMDs and healthy controls). After eliminating VOCs with low occurrence frequency, we obtained 174 ubiquitous VOCs and plotted ROC curves for everyone. Then 70 VOCs with $p < 0.05$ were selected for t-test, and 31 VOCs that have significant difference between two groups were selected shown in Table S1. To evaluate the discriminating performance of the selected biomarkers, we used 31 VOCs as independent variable to establish the diagnosis model by binary logistic regression analysis and the ROC curve was shown in Figure 2a. The datasets were randomly divided into two sets: training set (70% of the datasets) and validation set (30% of the datasets). Based on the established model, the sensitivity, specificity, and overall accuracy were 80.3%, 71.0% and 75.1%, respectively. Table 3 shows the discrimination results of the model for LC, PNMD and healthy control. The established model has high accuracy in distinguishing LC and healthy control. However, the model is unable to effectively discriminate between LC and PNMD due to the very low specificity, and the results are in accordance with previous study [9]. Due to the interference of PNMD patients, the diagnosis model should be optimized further for accurate LC diagnosis.

Diagnosis model optimization

Since the established diagnosis model is unable to effectively discriminate LC patients from PNMD patients, more variables that explicitly relate to LC should be introduced into the model to increase its discrimination capability. CT scan is commonly used for detecting changes in the lung parenchyma in clinical, and it was reported that the diagnostic accuracy of PET-CT for lung cancer was 93.5%, and the false positive rate was 6.5% [22] (the ROC curve of CT data only was shown in Figure 2b and the sensitivity, specificity, and overall accuracy were 85.8%, 91.1% and 88.8%, respectively). In addition, LC is closely associated with gender [23], age [24] and smoking status [25] of patients. Therefore, we established a new 35 variables diagnosis model by combining variables including CT, gender, age, smoking status and previous 31 VOCs through binary logistic regression analysis. Consequently, the sensitivity, specificity, and overall accuracy were significantly improved to 89.3%, 89.5% and 89.4%. The ROC curve of optimized LC diagnosis model is showed in Figure 2c and the AUC was 0.957 which means this model have a superb power for discriminating LC group and non-LC group.

Developing a sensor for monitoring exhaled breath that can replace GCMS is of great significance in clinical. However, simultaneous detection of multiple substances is a big challenge for sensor development and applications. Hence, reducing the input variables and maintaining high accuracy was imperative in the model

optimization. Previous 35 variables were used to establish random forests LC diagnosis model with 10 nodes and 600 decision trees. Then, top five variables with the threshold of MDA > 10 (Table 4) were selected to re-establish the diagnosis model. The ROC curve was shown in Figure 2d and the sensitivity, specificity, and overall accuracy of the model for LC diagnosis were 88.7%, 90.1% and 89.4%, respectively. As a result, the five variables LC diagnosis model established by RF has similar accuracy comparing with 35 variables model. Therefore, the final optimized model for LC diagnosis is obtained based on three selected VOCs, CT data and age. The largely reduced variables in this model enables convenient sensor development for clinical LC screening.

In Table 5, we summarized the discrimination accuracy of three lung cancer diagnosis models in LC vs. NLC. First, 31 variables model is the worst as the interference of PNMD patients. Then, 35 variables model has a superb power for discriminating LC group and non-LC group through adding LC relevant variables. At last, 5 variables model has similar accuracy.

The source of VOCs *in vivo*

VOCs detected in exhaled breath were generated by *in vivo* metabolism. Due to the insolubility of VOCs in the blood, the measurable VOCs can be reflected by breath exchange *via* the lungs. In Table 6, three selected VOCs above are divided into two groups and their main biological reaction processes as well as relative enzymes are presented.

Alkylbenzene is considered to be produced due to exogenous influences include smoking, alcohol and pollution. These exogenous substances leak into the cytoplasm and cause peroxidative damage to proteins, fatty acid and DNA. Most LC patients have a long smoking history, and toluene increased in the breath of smoking patients versus that in nonsmokers [26]. The defense mechanisms in the human body can eliminate exogenous substances by the cytochrome p450, glutathione S-transferases, sulfotransferases, and N-acetyltransferases enzyme system [27].

Alkane is generated by oxidative stress response of polyunsaturated fatty acids (PUFA) in cellular membranes. This process proceeds by a free radical chain reaction mechanism. In LC patients, cytochrome p450 enzymes which catalyze the oxidation of organic substances involve in the emission of VOCs *via* hydroxylation [28]. Pentane or ethane has been widely used as a sensitive and noninvasive indicator of lipid peroxidation *in vivo* [29].

Identification of DEGs in LC and enrichment analysis

To explore the changes of metabolism between LC patients and healthy controls, we compared gene expression levels between 347 LC samples and 22 normal samples from TCGA. After R statistical analysis (the code of R language can be found in supplementary materials), edgeR finally obtained 1159 DEGs. We set strict cut-offs with fold-change larger than 4 and FDR less than 0.001 to get highly significant genes which can distinguish the two groups. Table S2 lists the genes whose expression is significantly higher or lower

Group	N	Minimum	Maximum	Mean	Std
LC	197	230	436	355.7	37.8
PNMD	70	178	450	349.7	39.7
Healthy control	178	220	414	354.9	35.8

Table 2: VOC numbers in exhaled breath of LC, PNMD and healthy control groups

	LC vs. NLC ¹	LC vs. PNMD ¹	LC vs. HC ²
Sensitivity	80.20%	95.40%	93.40%
Specificity	71%	22.90%	81.50%
Overall accuracy	75.10%	76.40%	87.70%

¹Non-LC group

²Healthy control

Table 3: The discrimination accuracy of 31 VOCs based diagnosis model in LC vs. NLC, LC vs. PNMD and LC vs. HC

Variables	MDA
CT data	44.7
Age	22.1
620-14-4	16.8
526-73-8	13.9
108-87-2	11.0

Table 4: Top five variables with MDA>10 in RF model

in LC patients than in normal samples. The up-regulated genes were about twenty times larger than down-regulated genes, 1106 and 53 respectively. Figure S1 presents the Volcano Plot which shows the distribution about logFC (fold-change) and FDR of all genes with up-regulated genes (red) and down-regulated genes (green).

To understand the roles of DEGs playing *in vivo* of human body, we performed KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway and gene ontology (GO) terms enrichment analysis using DAVID to identify the biological process that enriches the DEGs. As a result, top ten ranked KEGG pathways were closely associated to lung cancer, especially alcoholism, nicotine addiction, metabolism of xenobiotics by cytochrome P450 (Table S3). Meanwhile, for the GO analysis, we categorized significant 30 terms in Figure 3. Obviously, these GO terms were significantly involved in cancer related general process such as DNA replication and repair, phosphorylation, and immune response.

Verification of VOCs through biological evidence

Kinds of genes encoded the key enzymes presented in Table 6, and those genes involved in the pathways which yielded specific VOCs were all differentially expressed in lung cancer samples (Table S2). For example, SULT4A1, GSTA8P, GSTA9P, NAA11 and NAALADL2-AS2 encode glutathione S-transferases, sulfotransferases, N-acetyltransferases, respectively, which play important roles in the production of alkylbenzene. These genes were all significantly up-regulated, and the logCPM were 4.6, 6.0, 5.6, 6.0 and 5.2 respectively. It means that the pathway of body's defense mechanism is more intense in LC patients, and excessive products would be released into the extracellular environment and enter the lung through blood circulation. This is the reason why exhaled breath from LC patients has more alkylbenzene such as 3-Ethyltoluene and 1,2,3-Trimethylbenzene.

Cytochrome p450 is a super enzyme family including various protein isoforms, and uses various molecules as substrates in

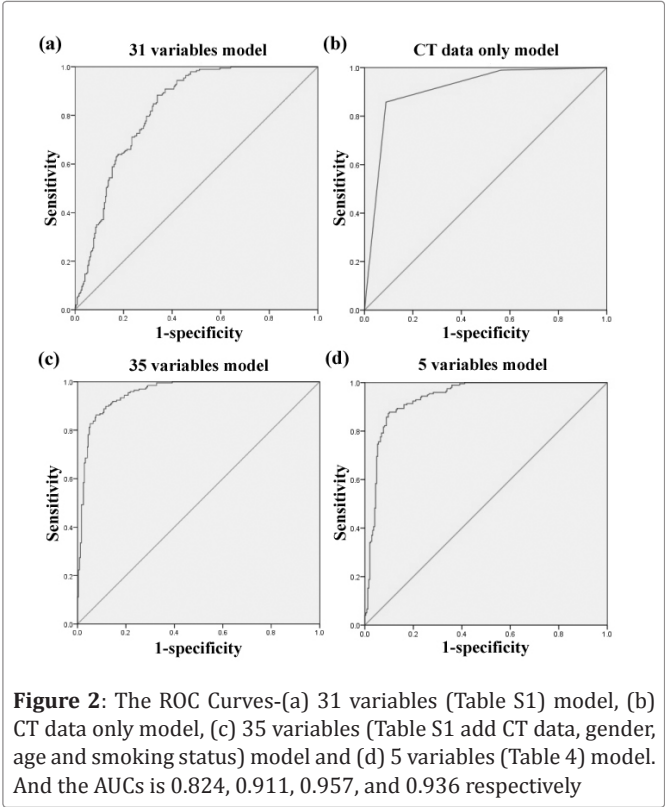
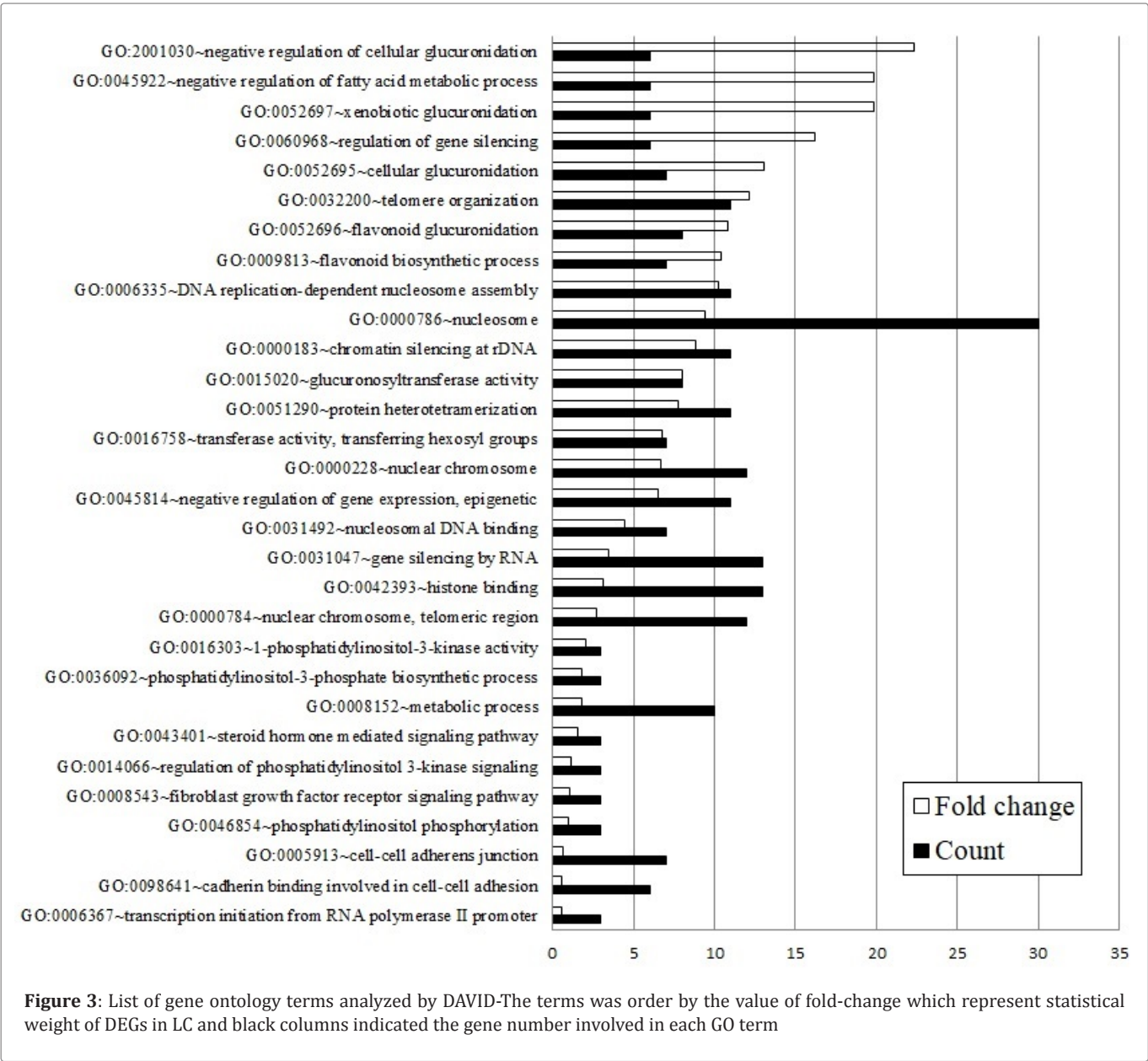


Figure 2: The ROC Curves-(a) 31 variables (Table S1) model, (b) CT data only model, (c) 35 variables (Table S1 add CT data, gender, age and smoking status) model and (d) 5 variables (Table 4) model. And the AUCs is 0.824, 0.911, 0.957, and 0.936 respectively

enzymatic reactions. This kind of enzyme is encoded by varies of genes like CYP1A2, CYP24A1, CYP26A1, CYP4F11, CYP4F3 and CYP2AB1P. According to our analysis results, CYP1A2 was down-regulated (logCPM=-4.4) and other five genes were up-regulated. It implies that this enzyme catalyzes different kinds of reactions involved in LC. Cytochrome p450 oxidizes a variety of structurally unrelated compounds, including steroids, fatty acids, and xenobiotics. For instance, CYP26A1 plays a key role in retinoic acid metabolism [30], CYP4F11 plays a key role in vitamin K catabolism by mediating omega-hydroxylation of vitamin K1 and K2 [31], and CYP4F3 catalyzes the omega-hydroxylation of leukotriene-B4 [32]. Importantly, alkane and alkylbenzene are the metabolites of these reactions. When CYP gene family was highly up-regulated, a mass of VOCs like methylcyclohexane would be generated apparently *in vivo*. For the down-regulated gene, CYP1A2 involves in the metabolism of aflatoxin B1 and acetaminophen [33]. Therefore, when CYP1A2 was down-regulated *in vivo*, it's reasonable to cause the retardation of related metabolism. In other words, aflatoxin B1 and acetaminophen would accumulate in the body or flow through other metabolic pathways which may generate the VOCs like 3-ethyltoluene and 1,2,3-trimethylbenzene as they all containing benzene ring. Consequently, alkylbenzene VOCs can be detected in the exhaled breath.

Conclusions

In conclusion, our work selected a set of VOCs from exhaled breath samples through GCMS. Besides, we established a LC diagnosis model based on 31 special VOCs and then improved sensitivity, specificity, and overall accuracy significantly to 89.3%, 89.5% and 89.4%, respectively, by adding LC relevant variables of gender, age, smoking status and CT data which can distinguish LC patients from non-LC people (including PNMDs and healthy controls). In order to develop sensor which can be widely used in clinical, we established an optimized five variables model with similar accuracy. Furthermore, to identify the source of VOCs, LC related metabolic pathways were obtained, and some pathways were consistent with biological process which generated VOCs *in vivo*. Overall, we established two optimized LC diagnosis model and illuminated the relationship between LC and VOCs.'



	31 variables model ¹	CT data only model	35 variables model ²	5 variables model ³
Sensitivity	80.2%	85.8%	89.3%	89.4%
Specificity	71%	91.1%	89.5%	90.1%
Overall accuracy	75.1%	88.8%	89.4%	89.4%

¹These 31 variables are showed in Table S1
²These 35 variables are Table S1 adding gender, age, smoking status and CT data
³These 5 variables are showed in Table 4

Table 5: The discrimination accuracy of different lung cancer diagnosis model in LC vs.NLC.

CAS	VOC name	Groups	Pathways	Relative enzyme
620-14-4	3-Ethyltoluene	Alkylbenzene	Body's defense mechanism	Cytochrome p450, glutathione S-transferases,
526-73-8	1,2,3-Trimethylbenzene			S-transferases, sulfotransferases, N-acetyltransferases
108-87-2	Methylcyclohexane	Alkane	Oxidative stress response	Cytochrome p450

Table 6: Pathways which generate the selected VOCs and its relevant enzymes

Acknowledgements

This research was supported by projects of Natural Science Foundation of China (No. 31571004, 31627801). And we thank all the volunteers for this study.

Conflict of Interest

The authors declare they have no conflict of interest.

Ethical approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. Ethical approval for human exhale breathes collection was obtained (The institutional ethics review committee of Sir Run Run Shaw Hospital, No. 20070525 and ChiCTR-DCD-15007106).

Informed consent

Informed consent was obtained from all individual participants included in the study.

References

- Parkin DM, Bray F, Ferlay J, Pisani P. Global cancer statistics, 2002. *CA Cancer J Clin*. 2005 Mar-Apr;55(2):74-108.
- Hu J, Qian GS, Bai CX. Chinese consensus on early diagnosis of primary lung cancer (2014 version). *Cancer*. 2015 Sep;121:3157-3164.
- Chen X, Xu F, Wang Y, Pan Y, Lu D, et al. A study of the volatile organic compounds exhaled by lung cancer cells in vitro for breath diagnosis. *Cancer*. 2007 Aug;110(4):835-844.
- Wang Y, Hu Y, Wang D, Yu K, Wang L, et al. The analysis of volatile organic compounds biomarkers for lung cancer in exhaled breath, tissues and cell lines. *Cancer Biomark*. 2012;11(4):129-137.
- Gordon S, Szidon J, Krotoszynski B, Gibbons R, O'Neill H. Volatile organic compounds in exhaled air from patients with lung cancer. *Clin Chem*. 1985 Aug;31(8):1278-1282.
- Sakumura Y, Koyama Y, Tokutake H, Hida T, Sato K, et al. Diagnosis by volatile organic compounds in exhaled breath from lung cancer patients using support vector machine algorithm. *Sensors*. 2017 Feb;17(2):287.
- Hakim M, Broza YY, Barash O, Peled N, Phillips M, et al. Volatile organic compounds of lung cancer and possible biochemical pathways. *Chem Rev*. 2012 Nov;112(11):5949-5966.
- Lourenço C, Turner C. Breath analysis in disease diagnosis: methodological considerations and applications. *Metabolites*. 2014 Jun;4(2):465-498.
- Zou Y, Zhang X, Chen X, Hu Y, Ying K, et al. Optimization of volatile markers of lung cancer to exclude interferences of non-malignant disease. *Cancer Biomark*. 2014;14(5):371-379.
- Balmain A, Gray J, Ponder B. The genetics and genomics of cancer. *Nat Genet*. 2003 Mar;33:238-244.
- Fidler MJ, Frankenberger C, Seto R, Lobato GC, Fhied CL, et al. Differential expression of circulating biomarkers of tumor phenotype and outcomes in previously treated non-small cell lung cancer patients receiving erlotinib vs. cytotoxic chemotherapy. *Oncotarget*. 2017 Apr;8(35):58108-58121.
- Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell*. 2011 Mar;144(5):646-674.
- Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, et al. Cancer genome landscapes. *Science*. 2013 Mar;339(6127):1546-1558.
- Lin EW, Karakasheva TA, Lee DJ, Lee JS, Long Q, et al. Comparative transcriptomes of adenocarcinomas and squamous cell carcinomas reveal molecular similarities that span classical anatomic boundaries. *PLoS Genet*. 2017 Aug;13(8):e1006938.
- Todd J, McGrath EE. Chest X-ray mass in a patient with lung cancer! *QJM: An International Journal of Medicine*. 2010 Sep;104(10):903-904.
- Andrews T, Wallace W. Diagnosis and staging of lung and pleural malignancy—an overview of tissue sampling techniques and the implications for pathological assessment. *Clinical Oncology*. 2009 Aug;21(6):451-463.
- Park SH, Goo JM, Jo CH. Receiver operating characteristic (ROC) curve: practical review for radiologists. *Korean J Radiol*. 2004 Jan-Mar;5(1):11-18.
- Maroof DA. Binary logistic regression. In: *Statistical methods in neuropsychology*. Boston (MA): Springer; 2012. P. 67-75.
- Breiman L. Random forests. *Machine learning*. 2001 Oct;45(1):5-32.
- McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res*. 2012 May;40(10):4288-4297.
- Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. 2009;4(1):44-57.
- Feng M, Yang X, Ma Q, He Y. Retrospective analysis for the false positive diagnosis of PET-CT scan in lung cancer patients. *Medicine*. 2017 Oct;96(42):e7415.
- Kiyohara C, Ohno Y. Sex differences in lung cancer susceptibility: a review. *Gend Med*. 2010 Oct;7(5):381-401.
- Tomasetti C, Li L, Vogelstein B. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science*. 2017 Mar;355(6331):1330-1334.
- Lee PN, Forey BA, Coombs KJ. Systematic review with meta-analysis of the epidemiological evidence in the 1900s relating smoking to lung cancer. *BMC Cancer*. 2012 Sep;12:385.
- Kischkel S, Miekisch W, Sawacki A, Straker EM, Trefz P, et al. Breath biomarkers for lung cancer detection and assessment of smoking related effects—confounding variables, influence of normalization and statistical algorithms. *Clin Chim Acta*. 2010 Nov;411(21-22):1637-1644.
- Guengerich FP, Shimada T. Oxidation of toxic and carcinogenic chemicals by human cytochrome P-450 enzymes. *Chem Res Toxicol*. 1991 Jul-Aug;4(4):391-407.
- Fontana E, Dansette P, Poli SM. Cytochrome p450 enzymes mechanism based inhibitors: common sub-structures and reactivity. *Cur Curr Drug Metab*. 2005 Oct;6(5):413-454.
- Terelius Y, Ingelman-Sundberg M. Metabolism of n-pentane by ethanol-inducible cytochrome P-450 in liver microsomes and reconstituted membranes. *The FEBS Journal*. 1986 Dec;161(2):303-308.
- White JA, Beckett-Jones B, Guo Y-D, Dilworth FJ, Bonasoro J, et al. cDNA cloning of human retinoic acid-metabolizing enzyme (hP450RAI) identifies a novel family of cytochromes P450 (CYP26). *J Biol Chem*. 1997 Jul;272(30):18538-18541.
- Edson KZ, Prasad B, Unadkat JD, Suhara Y, Okano T, et al. Cytochrome P450-dependent catabolism of vitamin K: ω -hydroxylation catalyzed by human CYP4F2 and CYP4F11. *Biochemistry*. 2013 Nov;52(46):8276-8285.
- Christmas P, Jones JP, Patten CJ, Rock DA, Zheng Y, et al. Alternative splicing determines the function of CYP4F3 by switching substrate specificity. *J Biol Chem*. 2001 Oct;276(41):38166-38172.
- Zhou H, Josephy PD, Kim D, Guengerich FP. Functional characterization of four allelic variants of human cytochrome P450 1A2. *Arch Biochem Biophys*. 2004 Feb;422(1):23-30.

Supplementary Information

No	CAS	Chemical formula	AUC	P
1	620-14-4	C ₉ H ₁₂	0.733	<0.001
2	526-73-8	C ₉ H ₁₂	0.726	<0.001
3	100-41-4	C ₈ H ₁₀	0.701	<0.001
4	1074-43-7	C ₁₀ H ₁₄	0.695	<0.001
5	496-11-7	C ₉ H ₁₀	0.668	<0.001
6	95-47-6	C ₈ H ₁₀	0.650	<0.001
7	66-25-1	C ₆ H ₁₂ O	0.643	<0.001
8	111-84-2	C ₉ H ₂₀	0.641	<0.001
9	469-61-4	C ₁₅ H ₂₄	0.640	<0.001
10	95-93-2	C ₁₀ H ₁₄	0.640	<0.001
11	109-60-4	C ₅ H ₁₀ O ₂	0.639	<0.001
12	824-22-6	C ₁₀ H ₁₂	0.619	<0.001
13	98-82-8	C ₉ H ₁₂	0.618	<0.001
14	108-87-2	C ₇ H ₁₄	0.617	<0.001
15	103-65-1	C ₉ H ₁₂	0.615	<0.001
16	95-16-9	C ₇ H ₅ NS	0.612	<0.001
17	465-24-7	C ₁₅ H ₂₆ O	0.610	<0.001
18	3075-84-1	C ₁₆ H ₁₈	0.597	<0.001
19	111-76-2	C ₆ H ₁₄ O ₂	0.597	0.001
20	108-90-7	C ₆ H ₅ Cl	0.590	0.001
21	108-10-1	C ₆ H ₁₂ O	0.587	0.002
22	124-19-6	C ₉ H ₁₈ O	0.586	0.002
23	80-56-8	C ₁₀ H ₁₆	0.585	0.002
24	3147-62-4	C ₇ H ₇ NO ₃	0.579	0.004
25	108-88-3	C ₇ H ₈	0.576	0.006
26	71-43-2	C ₆ H ₆	0.576	0.006
27	123-86-4	C ₆ H ₁₂ O ₂	0.575	0.006
28	6628-79-1	C ₆ H ₁₀ O ₃	0.573	0.008
29	124-13-0	C ₈ H ₁₆ O	0.570	0.011
30	546-28-1	C ₁₅ H ₂₄	0.565	0.019
31	13475-82-6	C ₁₂ H ₂₆	0.40	<0.001

Table S1: VOCs which had significant difference between LC group and Non-LC group

Symbol	log2 Fold Change	pval	FDR
AGER	-5.066191671	1.99E-50	7.12E-46
SFTPC	-4.760196879	1.54E-39	1.10E-35
GPD1	-4.368991814	1.80E-33	4.61E-30
MCEMP1	-4.478819866	6.17E-31	1.16E-27
GPM6A	-4.684060153	2.80E-27	3.71E-24
CA4	-4.659104021	1.08E-26	1.33E-23
VEGFD	-4.006876006	1.92E-23	1.86E-20
FAM107A	-4.931589542	4.41E-21	3.22E-18
FCN3	-4.547498784	6.14E-21	4.40E-18
CLDN18	-4.168121942	1.13E-19	6.96E-17
ITLN2	-4.564251093	2.18E-18	1.05E-15
GPIHBP1	-4.223253778	1.32E-17	5.91E-15
HBB	-4.719694054	2.69E-16	1.06E-13
HBA2	-4.318526926	3.35E-16	1.29E-13
SLC6A4	-5.690012452	1.32E-15	4.68E-13
SERTM1	-4.734513991	3.73E-15	1.15E-12
TMEM100	-4.422927328	4.97E-15	1.45E-12
CHIAP2	-4.271632148	4.61E-14	1.15E-11
GRIA1	-4.005399368	2.82E-13	5.98E-11
CD300LG	-4.904103699	7.99E-13	1.54E-10
GPA33	-4.113362348	1.10E-12	2.08E-10
RP11-88I21.2	-5.151310911	1.02E-11	1.60E-09
RP11-371A19.2	-4.980726918	1.34E-11	2.06E-09
ANGPT4	-4.421496001	2.45E-11	3.64E-09
RPL13AP17	-4.090112722	4.23E-11	5.95E-09
ALAS2	-4.636705086	6.43E-11	8.85E-09
FABP4	-4.066844491	2.15E-10	2.68E-08
TNR	-4.273779216	4.10E-09	3.78E-07
CD5L	-4.243498354	6.08E-09	5.35E-07
AP002856.5	-4.627135588	7.31E-09	6.28E-07
RP11-141J13.5	-4.324389421	1.33E-08	1.09E-06
KCNA4	-4.029872795	1.42E-08	1.15E-06
CLEC4M	-5.284055822	5.97E-08	4.25E-06
ADRA1A	-4.077761196	1.55E-07	9.76E-06
FREM3	-4.080046699	2.14E-07	1.29E-05
CSF3	-4.834555328	2.78E-07	1.64E-05
OR6K3	-4.669763137	2.85E-07	1.66E-05
RP11-287F9.2	-4.576179091	4.97E-07	2.69E-05
NCAPGP2	-4.521000117	5.24E-07	2.83E-05
HBA1	-4.961004043	6.04E-07	3.21E-05
DEFA3	-4.160715452	7.61E-07	3.95E-05
CTD-3214H19.16	-4.475503446	8.83E-07	4.50E-05
RP11-805I24.3	-4.093815158	1.21E-06	5.92E-05
PI16	-4.04635429	1.76E-06	8.26E-05
AC093787.1	-4.295785644	6.60E-06	0.000260743
PRG4	-4.746372047	1.80E-05	0.000604587
HBM	-4.103891706	2.09E-05	0.00069657
ITLN1	-8.183957217	2.53E-05	0.000824859

Table S2: The down-regulated genes in lung cancer

Pathways	Corrected <i>p</i> -Value	Relevant genes ¹
Alcoholism	1.82E-16	8360 8364 8367 51806 8368 8348 8329 8343 8342 8346 8345 2786 8366 8968 810 8332 158983 8331 8336 8335 8352 8339 8354 8355 8356 317772 8358 3018 85235 3013 2792
Systemic lupus erythematosus	1.11E-15	8360 8364 8367 8366 8368 8348 8329 8343 8342 8346 8345 8968 8332 158983 8331 8336 8335 8352 8339 8354 8355 8356 317772 8358 3018 85235 3013
Steroid hormone biosynthesis	1.81E-06	1544 1646 54659 10720 7366 54658 54575 54576 57577 54578 1109
Pentose and glucuronate interconversions	3.34E-06	57016 54659 10720 7366 54658 54575 54576 54577 54578
Retinol metabolism	4.61E-06	1544 54659 54576 7366 10720 54575 131 54577 54658 1592 54578
Ascorbate and aldarate metabolism	5.23E-06	54659 10720 7366 54658 54575 54576 54577 54578
Nicotine addiction	6.88E-06	2569 774 2556 2558 2890 2567 2566 55879 8973
Porphyrin and chlorophyll metabolism	9.49E-06	212 54659 10720 7366 54658 54575 54576 54577 54578
Metabolism of xenobiotics by cytochrome P450	1.17E-05	1544 22977 54659 10720 7366 54576 54575 131 54577 54658 54578
Drug metabolism - other enzymes	1.70E-05	7498 54659 10720 7366 54658 54575 54576 5577 54578

¹Relevant genes were represented by Entrez gene ID

Table S3: Dysregulated pathways in LC patients and its relevant DEGs

