Big cohort metabolomic profiling of serum for oral squamous cell carcinoma screening and diagnosis

Xihu Yang1,2 | Xiaowei Song3,4 | Xudong Yang1 | Wei Han1 | Yong Fu1 | Shuai Wang1 | Xiaoxin Zhang5 | Guowen Sun1 | Yong Lu1 | Zhiyong Wang1 | Yanhong Ni5 | Richard N. Zare3,4 | Qingang Hu1

1 Department of Oral and Maxillofacial Surgery, Nanjing Stomatological Hospital, Medical School of Nanjing University, Nanjing, Jiangsu, China
2 Department of Oral and Maxillofacial Surgery, Affiliated Hospital of Jiangsu University, Zhenjiang, Jiangsu, China
3 Department of Chemistry, Fudan University, Shanghai, China
4 Department of Chemistry, Stanford University, Stanford, California, USA
5 Central Laboratory of Stomatology, Nanjing Stomatological Hospital, Medical School of Nanjing University, Nanjing, Jiangsu, China

Correspondence
Richard N. Zare, Department of Chemistry, Fudan University, Shanghai, 200438, China.
Email: zare@stanford.edu
Qingang Hu, Department of Oral and Maxillofacial Surgery, Nanjing Stomatological Hospital, Medical School of Nanjing University, Nanjing, Jiangsu 210000, China.
Email: qghu@nju.edu.cn
Yanhong Ni, Central Laboratory of Stomatology, Nanjing Stomatological Hospital, Medical School of Nanjing University, Nanjing, Jiangsu 210000, China.
Email: Niyanhong12@163.com
X. Y. and X. S. contributed equally to this work.

Funding information
National Natural Science Foundation, Grant/Award Numbers: 21974027, 82173380; Jiangsu Province’s Key Provincial Youth Talents Program, Grant/Award Number: QNRC2016841; Nanjing Municipal Key Medical Laboratory Constructional Project Funding, Grant/Award Number: Since 2012; Nanjing Medical Science and Technique Development Foundation, Grant/Award Number: YKK18123; US Air Force Office of Scientific Research through Basic Research Initiative (AFOSR FA9550-16-1-0113) and the Multidisciplinary University Research Initiative (MURI) program, Grant/Award Number: (AFOSR FA9550-21-1-0170)

Abstract
The survival rate of oral squamous cell carcinoma (OSCC) can be greatly improved if intervention could be initiated as early as possible. This poses a technical demand for developing a sensitive screening and specific in vitro diagnosis method for OSCC. Herewith, a large cohort consisting of 241 healthy contrast (HC) and 578 OSCC patients were recruited for conducting the rapid metabolic profiling of trace volume serum using conductive polymer spray ionization mass spectrometry (CPSI-MS). Statistical analysis picked out 65 metabolite ions as potential characteristic markers for differentiating OSCC from HC. With the aid of a supporting vector machine (SVM), OSCC can be distinguished from HC with an accuracy of 98.0% by cross-validation in the discovery cohort and 89.6% accuracy in the validation cohort. Furthermore, orthogonal partial least-square-discriminant analysis (OPLS-DA) also initially showed the potential for OSCC staging, especially between T1/T2 and T3/T4 stages with an accuracy of 90.1%. CPSI-MS combined with SVM or OPLS-DA can not only quickly distinguish OSCC from HC but also predict the OSCC progression from T1/2 to T3/4 stages in a few minutes, making it a promising tool for both screening and diagnosing high-risk population.

Key points:
- Sixty-five characteristic metabolite ions significantly changed in OSCC serum metabolic profile compared to that in the HC group.
- CPSI-MS combined with SVM achieved 89.6% accuracy on the validation cohort for OSCC prediction.
- CPSI-MS/OPLS-DA can distinguish T1/T2 from T3/T4 stages with an accuracy of 90.1% by cross-validation.
INTRODUCTION

Oral cancer is one of the most common cancers in the head and neck region. There are around 377,713 new cases and 177,757 cases of death estimated worldwide in 2020 due to oral cancer. Oral squamous cell carcinoma (OSCC) contributes around 90% of oral cancers. Tobacco use (smoked or chewed), alcohol consumption, and human papillomavirus infection are regarded as high-risk factors for OSCC development. The diagnosis of OSCC includes a physical examination, radiography, computed tomography, magnetic resonance imaging, and histopathological examination of tissue biopsies. However, changes in molecular distribution at the primary carcinoma site are difficult to track at early stages before the histological lesion can be detected. In addition, there are still many cases not diagnosed until the advanced stage when distant metastases have happened, thereby missing the best opportunity for treatment. If a necessary intervention before tumorigenesis could be conducted, the currently maintained 60% 5-year survival rate is expected to be greatly improved.

Currently, tissue-based biopsy remains the gold standard in cancer diagnosis. It requires harvesting biospecimens by invasive procedures such as biopsies or needle aspirations. These procedures have common issues such as patient discomfort and sampling inaccuracy caused by tissue heterogeneity. By contrast, liquid biopsy has been increasingly considered as an alternative option for cancer detection because it can provide cancer-associated molecular information in a minimally invasive manner. Liquid biopsy is conducted by detecting tumor-associated markers in the circulating or excreted biological fluids such as saliva, urine, and serum. Currently, the detecting markers were primarily focused on exosomes, circulating tumor cells (CTCs), and circulating cell-free tumor DNA (cfDNA) that are shed into the bloodstream by cancer cells undergoing apoptosis or necrosis. Several DNA and mRNA species were reported to be associated with OSCC progresses, such as Gal-1, Gal-3, Transgelin, miR-24, miR-181, miR-196a, miR-10b, miR-18, lincRNA-p21, GAS5, and HOTAIR.

Gene- or protein-based clinical diagnosis mainly relies on the use of several immunoassays that introduce a hybrid probe or an antibody as a specific recognition element. This immune recognition-based multiplex detection is inevitably restricted by cross-reaction and spectral overlap in the readout. The analytical period and economic cost also increased with the introduction of more biorecognition probes. In contrast to gene and protein molecular detection, metabolomics-based in vitro diagnosis also has considerable promise because it provides the metabolic phenotype information that can not only precisely characterize the oncometabolite distribution at different stages but also help to guide the necessary therapy. Therefore, a highly sensitive and specific metabolomics-based approach is in urgent demand for preclinical screening and diagnosis among the high-risk population.

In the past decade, ambient ionization mass spectrometry (AIMS) has increasingly gained acceptance in the field of clinical diagnosis including cystic fibrosis, breast cancer, renal cell carcinoma, cervical cancer, and so on. Its most advantageous characteristic is the direct desorption and ionization of analytes from the sample matrix under atmospheric conditions. No or minor additional sample pretreatment is needed compared to conventionally used chromatography-mass spectrometry systems. Among various AIMS methods, paper spray ionization (PSI) is the leading one due to its wide availability in materials. PSI integrates well with the dried blood spot collection and storage with the following analytical procedures including solvent extraction, analyte desorption, and electric-field-induced spray ionization.

However, the plain paper substrate has drawbacks such as native impurities and strong retention of hydrophilic metabolites. Various surface coatings or modifications have been reported to overcome these issues. Our group also proposes using polymer substrate instead of paper and has achieved remarkable improvements in ionization efficiency, signal stability, and the wide coverage of hydrophilic species. This approach of polymer spray ionization has shown a good response for biological fluid metabolic profiling. In previous work, we have reported the practical value of conductive polymer spray ionization mass spectrometry (CPSI-MS) in the discrimination of OSCC with premalignant lesion (PML) and healthy contrast (HC). Combined with machine learning (ML) for high-dimensional data interpretation, this method can be performed with high accuracy at way less cost, trace sample consumption (<3 μL), and high speed (a few seconds per sample), making it a promising analytical tool for clinical assays. CPSI-MS/ML has shown its advantage in directly collecting hundreds of metabolites abundance information from a trace dried biofluid spot within a few seconds under atmospheric conditions, and in identifying key salivary metabolites and pathways involved in the progression from the PML to OSCC stage. The characteristic metabolites previously discovered in saliva were mainly narrowed to small molecules whose molecular weight is less than 500 Da.

Compared to saliva-based diagnosis, serum samples have advantages of a tightly controlled homeostatic environment and less external interference caused by diet. Serum is a more clinically available biofluid than saliva. It not only contains small metabolites but is also rich in lipid information. Thus, the use of serum can gain more lipid information, which can complement salivary metabolic profiling. In addition, analysis of serum allows us to judge the progression of OSCC, something that we failed to be able to do with saliva. The minimally invasive nature of blood-based samples, the wide distribution of metabolites, and evidence of changes in metabolites during OSCC initiation and progression, make blood-based metabolites attractive biomarker candidates. Currently, dozens of metabolites have been...
Diagram of the serum metabolic profiling workflow by CPSI-MS/ML. (a) Two cohorts of serum samples were collected from the OSCC and HC volunteers as the marker discovery and validation sets, respectively. (b) One drop of dried serum spot (3 μL) was loaded onto a conductive polymer tip. Once the extraction solvent was spiked, the high voltage was switched on to trigger the data acquisition. (c) The high-dimension metabolic profiles of different groups were classified and visualized under the constructed 3D features space by an unsupervised machine learning model; (d) From a statistical analysis, the discriminating metabolites were selected as features. (e) Given the data of the two cohorts as the training and test sets, a machine learning model was applied; (f) The serum metabolite markers were further validated at the tissue level and the combination was employed as the diagnostic panel. tSNE, t-distributed stochastic neighbor embedding; Ln(FDR), natural logarithm of false discovery rate; Lv, latent variable.

**MATERIALS AND METHODS**

**Volunteer recruitment**

Two cohorts of volunteers were recruited prior to formal surgery or chemotherapy during two separate periods. The first batch was recruited at the stage of serum metabolic profiling for metabolite marker discovery. It consisted of 100 healthy contrast (HC) volunteers and 154 OSCC patients which were recruited from February to April 2019. The second batch was employed for marker and prediction model validation. It consisted of 141 HC and 424 OSCC patients which were recruited from May to July 2019. All human subject research was conducted in compliance with the ethical guidelines established by the Nanjing Stomatological Hospital, Medical School of Nanjing University. OSCC cases have been formally diagnosed by histopathological examination whereas the HC volunteers are visiting patients who were diagnosed to be negative ones. The race, gender, ages, and body weight index were strictly matched between the two groups. More details
about patients’ clinical demographics can be seen in the supplementary information (Table S1).

Specimen collection and preparation

Overnight 12 h fasting is required before the intravenous blood sampling in the morning. The blood withdrawal volume is approximately 1 mL for generating 400 μL serum. The same brand of glass centrifuge tubes (BD Vacutainer) was used for blood collection. To avoid the metabolite changes before preprocessing, blood sample was temporarily stored at 4°C until natural coagulation. Serum was prepared by 2000 × g centrifugation for 10 min at 4°C after a blood clot was formed. All serum samples were saved under −80°C for long-term storage until use.

Metabolomic profiling by CPSI-MS

A full description of the CPSI-MS instrumentation can be found in a prior publication. After the serum was thawed under ambient conditions, 3 μL serum and 1 μL 4-chlorophenylalanine (internal standard, IS) were transferred onto the tip of a conductive polymer to form a dried serum spot for data acquisition. Upon addition of methanol-water (7:3, v/v, 3 μL) onto the dried serum spot for metabolites extraction, high voltage (+4.5 kV) was applied to the conductive polymer tip to trigger the spray ionization process. Data were recorded in the Department of Chemistry, Fudan University using an LTQ Orbitrap Velos mass spectrometer (Thermo Scientific, San Jose, CA, US.). The full scan range was set at m/z 50–1000 under positive mode. The data acquisition period for each sample lasts for 15 s. The intensity of each metabolite ion was normalized with the IS intensity of each sample ([M+Na]+, m/z 222.0292). Quality control (QC) samples were prepared by pooling equal volumes of 20 HC and 20 OSCC serum samples. QC samples were analyzed throughout the run to monitor the CPSI-MS system variation. OSCC and HC serum samples were run in alternation with the QC samples evenly inserted into the entire sequence every 30 samples.

Tissue validation by DESI-MSI

There were eight intact OSCC tissues resected during clinical surgery for the serum metabolite markers validation. For each tumor, a series of cryosections were sliced at a thickness of 12 μm and stored at −80°C until use. Each cryosection will first go through H&E staining and the histopathology check to delineate the cancer and normal region for reference. Another adjacent slice will go through the DESI-MSI experiment. A commercial DESI system (Prosolia, Indianapolis, U.S) was employed for tissue imaging. N, N-dimethyl formamide-acetonitrile (1:1, v/v) was used as the spray solvent with a flow rate of 1.0 μL/min and nebulizer gas pressure of 1.6 MPa. The impact angle between sprayer and section mounting stage was 56°. High voltage of +4.0 kV was provided by the mass spectrometer and applied onto the sprayer to generate the electrospray for desorbing and ionizing the components across the tumor cryosection (12 μm). Target ion image reconstruction was achieved using Massimager (Chemmind Technologies Co., Ltd, China).

Metabolomics data processing

All raw files were first converted into cdf format by the Xcalibur (Thermo Fisher Scientific, San Jose, CA, US.) and then imported into MATLAB 2020a (Mathworks, Natick, MA.) for batch data preprocessing using the self-programmed script. Each sample’s metabolomic profile was presented by averaging the mass spectra over 10 continuous scans in the corresponding time window. The mass tolerance of each bin was set at ± 0.005 m/z when extracting the peak intensity information. Those peaks in the average mass spectrum which have intensity lower than 500 are regarded as background noise and discarded. A bin was treated as a missing value if it failed to be detected from a sample. No missing value imputation was made to avoid artificial statistical results in the univariate analysis. To reduce the matrix data volume, the bin that possesses more than 50% missing values among the first cohort of 254 samples was discarded. Finally, there were 1518 bins initially extracted to characterize the metabolomic profile. A data matrix was constructed with each row representing one case and each column representing one bin variable. Then, the matrix goes through the IS normalization, natural log transform, zero-centering, and unit variance scaling before applying univariate analysis, multivariate analysis, and machine learning modeling. The data processing was done at Fudan University and Stanford University.

Statistical analysis

The unsupervised metabolic profile differentiation between OSCC and HC groups was first conducted with the t-stochastic neighbor embedding (t-SNE) in the MATLAB program. Rank sum test was first implemented separately among the two cohorts to search the OSCC and HC groups for significantly changed metabolite ions. The false discovery rate (FDR) was estimated with Benjamini and Hochberg method to adjust P-value and assess the statistical significance. The ion will be selected if its FDR value is lower than 0.05. Only ions that are significantly changed both in the discovery cohort and validation cohort will be regarded as potential serum metabolite markers. Finally, a metabolite with a fold change larger than 1.5 or smaller than 0.67 will be included for further validation at the tissue section by DESI-MSI. Orthogonal partial least squares discriminant analysis (OPLS-DA) was used for OSCC staging by aid of SIMCA-P (Umetrics, Umea, Sweden). Variables with importance in projection (VIP) values higher than 1.5 were considered to contribute strongly to the pattern recognition of different OSCC stages. Prism (GraphPad Software, USA) was employed for preparing box plots, heatmaps, and receiver operating characteristic (ROC) curves.
Machine learning modeling

Two cohorts of OSCC and HC serum cases were recruited for the machine learning model development. For the OSCC screening modeling, the first cohort (100 HC + 154 OSCC) was used for classification model comparison and training. The 5-fold cross-validation was conducted in the first cohort to assess the model training performance. The MATLAB in-built APP “classification learner” was employed to select the optimal model for training and validation. A variety of classification models were investigated including linear discriminant analysis (LDA), supporting vector machine (SVM), k-nearest neighbor (KNN), and ensemble method. A confusion matrix was used to display the classification results and calculate the general accuracy, sensitivity, and specificity. The F1 score was used as the single metric to assess different models’ fitting performance. Finally, the second cohort (141 HC and 424 OSCC) was used as the validation set. The area under the curve (AUC), specificity, and sensitivity were used as the metrics for comparing different machine learning models’ generalization ability to give a fair assessment of the pretrained model performance on the unseen data. For the OSCC staging study, the two cohorts of OSCC cases were combined to obtain sufficient samples for each stage (T1, n = 139; T2, n = 167; T3, n = 128; T4, n = 144). Then the fivefold cross-validation was used for evaluating the prediction accuracy.

Analyte annotation and marker identification

Tables S2 and S3 present the annotated metabolite ions discussed in this study. The metabolite ions of interest were first searched through HMDB (http://hmdb.ca) and Metlin (https://metlin.scripps.edu) with the 5.0 ppm mass tolerance. The type of adduct ions included [M+H]⁺, [M+Na]⁺, [M+K]⁺, [M-H₂O+H]⁺, [M+2Na-H]⁺, [M+2K-H]⁺, and [M+NH₄]⁺. Only those candidates with a reported presence in humans were given consideration. For those unknown significantly changed ions, MS/MS experiments were performed to match the collision-induced dissociation (CID) fragmentation patterns either with given standards or recorded MS/MS spectra in HMDB and Metlin.

RESULTS

Serum metabolic profiling of OSCC

Collected 819 serum samples consisting of 241 HC and 578 OSCC were divided into a development cohort and a validation cohort for serum marker discovery and confirmation. There were 367 ions selected to characterize the global metabolic profiles of HC and OSCC. Figure 1a presents the mass spectra averaged from 578 OSCC and 241 HC groups. From the average mass spectra, it can be clearly observed that the peaks at m/z 135.0029 (lactic acid, [M+2Na-H]⁺), 203.0527 (glucose, [M+Na]⁺), 304.2611 (oleamide, [M+Na]⁺) were elevated in the OSCC group compared to the HC group. More discriminative fingerprint peaks had to be found by statistical tests, which are described later.

To visualize the difference between HC and OSCC metabolite patterns, an unsupervised machine learning method, t-SNE, was introduced to reduce the high-dimensional metabolite ions information into a three-dimensional (3D) feature space. In the constructed 3D feature
FIGURE 2  Discovery and validation of the serum metabolite markers. (a) Venn graph displayed the number of discriminating metabolite ions selected in the development and validation cohort and the number of common ions; (b) The most significantly metabolite ions with the largest fold changes were highlighted by a volcano graph; (c) The top 10 metabolites’ relative expression levels were visualized by a heatmap; (d) Two typical serum lipids relative intensities in serum displayed by box plots. (e) Their spatial distribution in the intact OSCC tissues were visualized and compared by DESI-MSI under guidance (f) of H&E staining images. The scale bars in (e) and (f) denote 1.0 mm whereas the scale bar zoomed in the H&E staining image denotes 50 μm. Ln(FDR), the natural logarithm of false discovery rate

space, serum cases from the same group were well clustered whereas those cases from different groups can be separated (Figure 1b). This result demonstrated there exists a substantial difference in serum metabolic profiles that can be used for OSCC and HC prediction.

Discovery and validation of serum metabolite markers

The rank sum test was employed to search for low abundance discriminating ions. In the development cohort, there were 241 significantly changed ions in OSCC compared to the HC (FDR < 0.05). When the same procedure was conducted in the validation cohort, 218 ions were found to have significant differences. After overlapping the two batches of discriminating ions, 65 ions were confirmed to not only have statistical significance but also to become upregulated or downregulated in the same direction (Figure 2a). The non-overlapped metabolites may come from several sources such as inter-batch variation in individuals and sample storage. After removing ions that were either redundant or failed to meet the fold change criteria (larger than 1.5 or smaller than 0.67), 39 metabolites were finally selected as potential characteristic marker candidates (Table S4). A volcano graph highlighted the top 10 metabolites with the most obvious fold changes (Figure 2b), which shows that lipid molecules are the predominant species in the serum including glycerophosphocholine (GPC), lyso-glycerophosphocholine (Lyso-GPC), acyl carnitine, diacylglycerol (DG), sphingolipids (e.g., sphingosine-1-phosphate). Figure 2c shows the relative expressions of these top 10 metabolites across the development and validation cohorts in the form of a heatmap.

In situ validation of the serum metabolite markers

The top 10 characteristic metabolites discovered in serum were further analyzed at the tissue level. Frozen cryosections of eight intact OSCC tumor tissues were prepared for DESI-MSI analysis. Then the
spatial distribution of target metabolites across the resected OSCC tissues was mapped and compared. The in situ validation by DESI-MSI revealed that 8 of top 10 metabolites had the same trends with OSCC in tissue and serum (Figure S1). Taking lysoPC(20:4) and PC(34:6) as examples, the former one declined in the cancer region compared to the negative contrast region whereas the latter one was elevated in the cancer region and can specifically delineate the cancer margin. The low expression of lysoPC(20:4) and high expression of PC(34:6) in the cancer region were consistent with their trends in the serum cohort analysis (Figure 2d-f). As for the remaining metabolites, lysoPE(20:3), and sphingosine-1-phosphate, the former one had an opposite trend (up-regulated) whereas the latter one was not so obvious in the increased level. We propose that was largely caused by the individual variance and limiter tissue number in this study. The large-scale tissue validation should be conducted in a further study.

Expression of salivary metabolite markers in serum

Given the 106 characteristic metabolites previously studied in salivary metabolic profiling,30 the extent of their changes in serum between the OSCC and HC group were investigated. For this inter-specimen validation purpose only, the serum samples from two cohorts were combined to implement the rank-sum test. As a result, 52 out of 106 metabolites discovered in the previous saliva metabolomics were found to remain at abnormal levels in serum (Table S5), although changes in serum for most of these metabolites were not as obvious as those in saliva. Moreover, 33 out of these 52 metabolites in serum showed the same change trends as those in saliva (OSCC vs HC). Altogether, there were 65 metabolites discovered to be changed in OSCC compared to HC with statistical significance (FDR < 0.05). These metabolites were treated as serum marker candidates that have the potential discrimination power for OSCC prediction model development.

Feature selection and machine learning model development

A variety of classification models were trained to determine the most suitable one for further development. At the initial stage, all the selected metabolites in the univariate analysis were included as feature sets to train models. As a result, although all models can achieve perfect performance in the first cohort of 254 cases with an accuracy of no less than 90% (Table S6), their performances in the second cohort (as the unseen cases) differed from one to another. The SVM achieved the general accuracy at 86% with the maximum area under curve (AUC) value at 0.86 (95% CI: 0.82-0.90). From the receiver operating characteristic (ROC) curves, SVM also gains the highest diagnostic performance with a sensitivity (85.1%) and specificity (90.6%) (Figure 3a). Therefore, SVM was selected as the optimal model for further tuning.

Feature selection is a critical step to avoid overfitting by reducing the model complexity. Recalling that all metabolite ions have statistical significance between the two groups, there were various possibilities for feature selection and combination for model development. To achieve a more robust machine learning model, it is necessary to select the optimal set of metabolites as the characteristic. For this purpose, we choose a wrapper-type feature selection strategy that evaluates the chosen machine learning model’s performance after training with different candidate feature subsets.43 Briefly, the absolute weights of the 65 metabolite ions in the initial SVM model were ranked to evaluate their discriminating powers. Then the training sets with features consisting of the top 60, 50, 40, 30, 20, 15, 10, 5, and 2 metabolite ions were composed and trained in the first cohort. As is shown in Figure 3b, the SVM model’s performance with different feature subsets maintained stable behavior for the training set whereas the accuracy greatly dropped in the test set when the numbers of features were less than 15.

The relative expression levels of these 15 metabolite ions in the test set are shown in Figure 3c and Table S7. These 15 metabolite ions also had statistical significance no matter in the development cohort or validation cohort, proving their feasibility as the clinical markers. The classification result on the test set was visualized in the dimension-reduced space composed of the first two principal components, in which it is seen that more than 500 HC and OSCC samples can be ideally separated (Figure 3d). The confusion matrix showed that the optimal SVM model can obtain a sensitivity of 92.1% and specificity of 81.1% for OSCC detection (Figure 3e). The final prediction accuracy reached 89.6% on the test set (Table S8).

Serum metabolomic profiling for OSCC stages

We also investigated whether serum metabolomic pattern differences exist not only between HC and OSCC but also among different stages (from stage T1 to stage T4). The OPLS-DA model visualized the distribution of the OSCC cases staged from T1 to T4 at two cohorts (Figure 4a). It can be seen that the OSCC at T1 and T2 stages can be ideally separated from these samples at the T3 and T4 stages. After combining T1 with T2 and T3 with T4, the OPLS-DA model performance can be greatly improved to the accuracy of 90.1% (CV = 5). Unfortunately, there was no obvious separation between T1 and T2 or T3 and T4. The variables that made a high contribution to this T1/T2 and T3/T4 separation were searched according to their variable importance on projection (VIP) values. The variables with VIPs large than 1.5 were picked (Table S9). After removing redundant ions, the top 50 metabolites were annotated and their relative contents in serum were displayed in a heatmap. It was worth noting that four of the top 10 metabolites were discovered in the univariate analysis (DG(34:2), oleamide, lysoPC(20:4), and nonanoyl carnitine) also contributed to this T1/T2 and T3/T4 stages discrimination (Figure 4b). As shown in Figure 4c, the relative contents of these metabolites in the OSCC group (T1-T4) were completely different from those in the HC group. Furthermore, they also showed increasing or decreasing trends from T1 to T4. Unfortunately, none of these top 50 metabolites showed statistical significance among the four stages by the analysis of one-way variance (ANOVA).
FIGURE 3 The development of serum metabolomics-based machine learning model for OSCC diagnosis. (a) Different machine learning models were initially investigated by comparing their diagnostic performance on the test set. SVM was chosen as the optimal one; (b) The number of features was investigated by a sequential feature selection strategy. Fifteen features were sufficient for the SVM model to achieve the optimal predicting accuracy on the test set. (c) The relative fold changes of these 15 metabolite ions on the test set (OSCC vs HC) were visualized; (d) The distribution of two cohorts of HC and OSCC cases and their decision boundary given by an SVM were displayed in the feature space constructed with the first two principal components; (e) Confusion matrix given by an SVM model with finally selected 15 metabolite ions as the input features.

DISCUSSION

This study has demonstrated several advantages of CPSI-MS/ML for OSCC diagnosis from serum samples. From the aspect of data collection efficiency, CPSI-MS realized quick collection of high-dimension metabolomic data from each case directly with a timescale of seconds. The total analytical period for these two cohorts of 819 serum samples took only 12 h, which satisfies practical requirements for clinical screening. CPSI-MS is quite suitable for the rapid, direct metabolomic profiling from a dried spot of biological fluid such as saliva, serum, or even whole blood. A basic methodology investigation was conducted in this study. A series of serum samples were evenly distributed among the whole test sequence. Then, the variations of the first two principal components (PC1-PC2) were analyzed. The relative standard deviations (RSD values) of PC1 and PC2 fall into the acceptable levels at 18.7 % and 31.2 % (Figure S2), respectively, meeting the basic requirement of qualitative analysis.44 This result is largely because data acquisition from the whole cohort can be completed in one working day. The short period of single case analysis by CPSI-MS could make the large cohort assay conducted more effectively. The number of QC samples introduced for monitoring and normalizing the MS system variation was also reduced. This variation is a critical factor that cannot be ignored, especially compared to data taken from traditional LC-MS or GC-MS systems. With aid of a pre-trained machine learning model, the high-dimension metabolome data can be transferred into accurate diagnostic information almost instantly without biased interpretation by practitioners, facilitating its practical value in precision medicine. Additionally, we should point out the pattern recognition in the PCA or t-SNE clustering heavily depends on the features included. In Figure S2, the discovery cohort and validation cohort were distributed to different clusters because in this stage, all mass bins in the general metabolic profile were included as features. Despite this interbatch shift, the final SVM model still achieved a good predictive performance on both datasets. This result further indicated that the features we selected for the SVM model were sufficient to have a robust response.

From the studies of serum metabolomics reported here and the previous saliva metabolomics, the OSCC-associated discriminating metabolites were identified, respectively. Given the identified metabolites, a pathway enrichment analysis was implemented to locate those metabolic pathways that are influenced in serum and saliva (Table S10) by searching in the open-source platform Metaboanalyst.45 A
The hypergeometric test was employed to evaluate the statistical significance of a pathway. The relative-betweenness centrality was used as the metric in topology analysis to evaluate the pathway impact, which ranges from zero to one. As a result, the four representative metabolic pathways (histidine metabolism, arginine biosynthesis, arginine, and proline metabolism, aminoacyl-tRNA biosynthesis) discovered in the saliva remained highlighted in the serum level, whereas their impact or significance did not rank at the top. Instead, lipids-related metabolism becomes the major pathway including glycerolipid (GL), glycerophospholipid (GPL), and sphingomyelin (SM) (Figure 5). According to the fold changes of these metabolites (Tables S4 and S5), the changes of many metabolites become less obvious in serum, although the 52 discriminating metabolites discovered in the saliva study still had abnormal abundance in serum. This was observed mostly among the metabolites located in the histidine, arginine, and proline metabolic pathways, which were the major changed pathways in the saliva of the OSCC group. In contrast, the GL, GPL, and SM molecules in serum become the major discriminating markers (Figure S3).
Because OSCC occurs in the oral cavity, cancer cells might scavenge nutrient supply either endogenously from the local blood circulation or exogenously from the excretion of the salivary gland. In turn, the OSCC cells’ metabolic products will also be exchanged with the extracellular environment and transported through the circulation system. Therefore, this inter-specimen derived difference in dysregulated metabolic pathways might be attributed to the complex biomass transport and exchange differences between the oral environment and endogenous circulation environment. Another possibility for explaining why salivary discriminating metabolites have diminished significance in serum might be caused by dilution in the global blood circulation. This suggests the possible value of employing serum metabolome data complementary with the salivary metabolome data for OSCC diagnosis based on serum lipid features.

It is known that cancer cells can utilize massive nutrients to support their uncontrolled proliferation. Carbohydrates, amino acids, nucleotides as well as fatty acids were all their target biomass not only as the basic building blocks for proteins, glycans, nucleic acids, and bilayer lipids of membranes but also as the functional agents such as energy fuels, signaling factors, and transport intermediates. The excessive energy consumption was also observed by the abnormal levels of glucose, lactic acid, free fatty acids (e.g., palmitic acid, palmitoleic acid, caprylic acid, linolenic acid), mono-acyl glycerol (e.g., MG(14:0), MG(16:1), and MG(16:0)), and acylcarnitine (e.g., acetylcarnitine, nonanoyl carnitine, 6-keto-decanoylcarnitine, and pentadecenoyl carnitine) for glycolysis and β-oxidation in the mitochondrion. GL, GPL, and SPL are not only the critical constituents for building the bilayer membrane systems but also served as the regulators for signaling. The abnormal SPL metabolism (e.g., sphingosine 1-phosphate, sphinganine, and phytosphingosine) suggests cell proliferation dysregulation. These dysregulated metabolite markers could not only serve as potential markers for OSCC diagnosis but also might assist roughly evaluating the OSCC stages.

Serum metabolome-based profiling and metabolites panel-based detection can serve as the molecular diagnosis approach complementary with the traditional tissue-based histopathology and routine visual examination. More than half of serum discriminating metabolites can be traced to the primary lesion site of OSCC tissue by the DESI-MSI confirmation, proving the feasibility of using serum metabolome information for OSCC detection. As for the rest of the discriminating metabolites that failed in DESI-MSI detection, it might be attributed to the limited sensitivity of DESI-MSI especially in these species with low abundance and ionization efficiency. This possible explanation needs to be confirmed in the future with the aid of an LC-MS or GC-MS system after tissue extraction and the analyte enrichment process.

It was also found that the serum metabolome has the potential for roughly assessing OSCC stages (Figure 4a). Although so far, no inter-stage statistical significance was found among any single serum metabolite (Figure 4b), a clear pattern difference appears especially when the OSCC stage was developed from T1 or T2 into T3 or T4 (Figure 4c). This result also emphasizes the fact that a single metabolite signature is neither specific nor sensitive enough to indicate the OSCC occurrence and development compared to the combination of characteristic metabolites in the form of a diagnostic panel. We must admit that the features selected for the OSCC diagnosis only achieved the distinction between two staging groups, (T1, T2) and (T3, T4). The criteria of traditional hypothetical tests in univariate analysis (conventionally referred to P or FDR < 0.05) might be too cautious for OSCC staging to pick out important features in profiling-based prediction methods. In another aspect, it also implies the importance of multivariate analytical method or even machine learning method in discerning these critical feature variables for the profile pattern recognition.

CONCLUSION

We find that the serum metabolic profile obtained by CPSI-MS and analyzed using machine learning can reflect oral cancer development. Most discovered significant metabolites in serum were also found in saliva and cancer tissue, demonstrating the potential of serum for in vitro molecular diagnosis of OSCC. By cohort analysis using CPSI-MS, we found that histidine metabolism, arginine and proline metabolism, sphingolipid metabolism, and aminoacyl-tRNA biosynthesis were present in serum. These findings provide potential clinical markers for indicating OSCC tumorigenesis. We have demonstrated that CPSI-MS is a promising ambient ionization mass spectrometry tool that offers cost-effective performance in monitoring hundreds of biofluidic metabolites only with minor sample pretreatment. The combination of CPSI-MS with ML enabled excellent OSCC prediction performance.
(89.6 % accuracy). More surprisingly, the CPSI-MS combined with an OPLS-DA model could well differentiate the (T1, T2) with (T3, T4) stages (90.1% accuracy by cross-validation). All these findings indicate that CPSI-MS/ML can be a very useful tool to provide a simple, fast, affordable method both for OSCC screening and diagnosis.

**AUTHOR CONTRIBUTIONS**


**DATA AVAILABILITY STATEMENT**

Serum metabolomic profile data of all study cases are available in open access through visiting the Open Science Framework (OSF, https://osf.io/wnuf/).

**TRANSPARENT PEER REVIEW**

The peer review history for this article is available at https://publons.com/publon/10.1002/ntls.20210071.

**ETHICAL STATEMENT**

Human sera and tissue samples were collected in strict observance of the ethical code of Nanjing Stomatological Hospital, Medical School of Nanjing University. All patients gave written consent.

**CONFLICT OF INTEREST**

The authors declare no potential conflicts of interest.

**ORCID**

Richard N. Zare https://orcid.org/0000-0001-5266-4253

**REFERENCES**


