Metabolomic Profiling of Serum for Large Cohort Oral Squamous Cell Carcinoma Diagnosis

Xihu Yang[1], Xiaowei Song[2], Xudong Yang[1], Wei Han[1], Yong Fu[1], Shuai Wang[1], Xiaoxin Zhang[1], Guowen Sun[3], Yong Lu[1], Zhiyong Wang[1], Yanhong Ni[1], Richard Zare[4], and Qingang Hu[1]

[1]Medical School of Nanjing University
[2]Fudan University
[3]Nanjing University Medical School Affiliated Nanjing Drum Tower Hospital
[4]Stanford University

September 27, 2021

Abstract

Background: Oral squamous cell carcinoma (OSCC) accounts for 90 % of oral cancers. If a necessary intervention before tumorigenesis could be conducted, the current 60% 5-year survival rate would be expected to be majorly improved. This fact motivates the search for developing a highly sensitive and specific in vitro diagnostic method to conduct rapid OSCC screening. Method: Serum samples from 819 volunteers, consisted of 241 healthy contrast (HC) and 578 OSCC patients, were collected, and their metabolic profiles were acquired using conductive polymer spray ionization mass spectrometry (CPSI-MS). Univariate analysis was used to select significantly changed metabolite ions in the OSCC group compared to the HC group. Identities of these metabolite ions were determined by MS/MS experiments and reconfirmed at the tissue level by desorption electrospray ionization mass spectrometry (DESI-MS). The supporting vector machine (SVM) algorithm was employed as the machine learning model to implement the automatic prediction of OSCC. Results: Through statistical analysis, 65 metabolites were selected as potential characteristic marker candidates for serum OSCC screening. In situ validation by DESI-MSI revealed that 8 out of top 10 metabolites showed the same trends of change in tissue and serum. With the aid of machine learning, OSCC can be distinguished from HC with an accuracy of 98.0 % by cross-validation in the discovery cohort and 89.2% accuracy in the validation cohort. Furthermore, orthogonal partial least square-discriminant analysis (OPLS-DA) also showed the potential for recognizing OSCC stages. Conclusion: Using CPSI-MS combined with SVM, it is possible to distinguish OSCC from HC in a few minutes with high specificity and sensitivity, making this rapid diagnostic procedure a promising approach for high-risk population screening.

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.[1] 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.[2]The diagnosis of OSCC includes a physical examination, radiography, computed tomography, magnetic resonance imaging, and histopathological examination of tissue biopsies.[3, 4] However, changes in molecular distribution at the primary carcinoma site are difficult to track at early stages before the histological lesion can be detected.[5] 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 current 60% 5-year survival rate is expected to be majorly improved.[4]

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 exosome, circulating tumor cells (CTCs), and circulating cell-free tumor DNA (cfDNA) which are shed into the bloodstream by cancer cells undergoing apoptosis or necrosis. Several DNA and mRNA species were reported to be associated with OSCC progressions, such as Gal-1, Gal-3, Transgelin, miR-24, miR-181, miR-196a, miR-10b, miR-18, lincRNA-p21, GAS5, and HOTAIR.[6-13]

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.

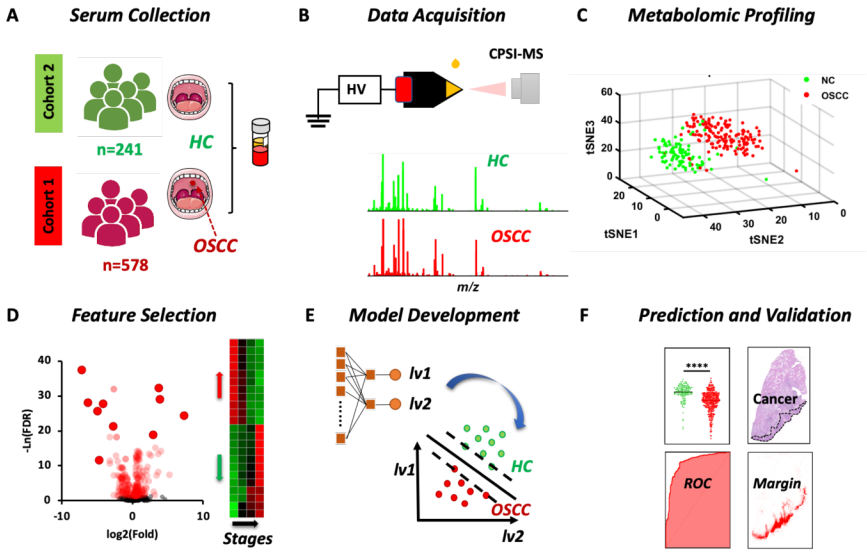
In contrast to gene and protein molecular detection, metabolomics-based*in vitro* diagnosis also has a considerable promise because it provides the metabolic phenotype information that can not only precisely characterize the oncometabolite distribution at different stages but also help to guide the necessary therapy.[14] Therefore, a highly sensitive and specific metabolomics-based approach is in urgent demand for preclinical screening among the high-risk population.

In recent years, ambient ionization mass spectrometry has gradually gained interest in the field of clinical diagnosis owing to its advantages in free from laborious pretreatment, wide coverage of metabolite species, and high-throughput metabolome information monitoring among various biological samples. Combined with machine learning for high-dimension data interpretation, it can be performed with comparable accuracy at way less cost. [15-20]

In previous work we have reported the practical value of conductive polymer spray ionization mass spectrometry combined with machine learning (CPSI-MS/ML) in the discrimination of OSCC with premalignant lesion (PML) and healthy contrast (HC).[21] 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,[22] 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. Serum is a more clinically available biofluid that not only contains small metabolites but is also rich in lipid information. The minimally invasive nature of blood-based samples, the sensitive feature of metabolites, and evidence of changes in metabolites during OSCC initiation and progression, make blood-based metabolites attractive biomarker candidates.[23, 24] Currently, dozens of metabolites have been reported to be dysregulated with OSCC malignant progression, including ketones, malonate, glutamine, propionate, valine, tyrosine, serine, methionine, and choline. [25-30]

Given the hypothesis that the serum probably contains more OSCC-associated metabolic phenotype information, there were two concerns that needed to be investigated in this study: (1) whether the previously discovered salivary metabolites can still be significantly different among HC and OSCC in the serum to serve for preclinical screening; and (2) whether the significantly different metabolites in the serum can be not only used for discriminating OSCC from HC but also for discerning OSCC at different stages (T1, T2, T3, T4). Therefore, the aim of this study was to develop panels of serum metabolite markers for OSCC screening. The potential of serum metabolic profiling for staging was also preliminarily investigated. With the aid of the CPSI-MS/ML approach, we believe that a low-invasive serum diagnosis can be realized to provide a quick, accurate, cost-effective diagnosis of OSCC.**Scheme 1** describes the general workflow that is followed.



Scheme 1. 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 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.

Materials and Methods

Volunteers Recruitment

Two cohorts of volunteers were selected, a discovery group (254), and a validation group (565), both containing individuals who have been diagnosed with OSCC prior to surgery or chemotherapy and individuals diagnosed no OSCC, called healthy contrast (HC). Specifically, 154 individuals have OSCC and 100 are HC in the discovery group, and 424 have OSCC and 141 are HC in the validation group. All human subject research was conducted in compliance with the ethical guidelines established by the Nanjing Stomatological Hospital, Medical School of Nanjing University. The race, gender, ages, and body weight index were strictly matched between the two groups. More details about patients' clinical demographics can be found in the supplementary information (**Table S1**).

Specimen Collection and Preparation

Overnight 12 hours' fasting is required before intravenous blood sampling in the morning. The blood withdrawal volume is approximate 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 samples were temporarily stored at 4 until natural coagulation occurred. Serum was prepared by 2,000g centrifugation for 10 minutes at 4 after blood clotting. All serum samples were saved under -80 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.[21,22] After the serum

was thaw under ambient conditions, 3 μL serum and 1 μL 4-chlorophenylanine (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 seconds. The intensity of each metabolite ion was normalized with the IS intensity of each sample ($[M+Na]^+$, m/z222.0292). 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 alteration with the QC samples evenly inserted into the entire sequence every 30 samples.

Tissue Validation by DESI-MSI

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 at 1.0 μL/min and nebulizer gas pressure at 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, US.) 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. There were 1518 peaks initially extracted to characterize the metabolomic profile. A data matrix was constructed with each row representing one case and each column representing one peak variable. To reduce the matrix data volume, the peaks that possessing more than 50 % missing values among the first cohort of 254 samples were discarded. No missing value imputation was conducted to avoid artifact statistical results in univariate analysis.[31]Then, the matrix goes through the IS normalization, natural log transform, zero-centering, and unit variance scaling before univariate analysis, multivariate analysis, and machine learning modelling is applied. 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.[32] 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 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 modelling, 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), logistics regression, decision tree (DT), naïve Bayesian (NB), 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, true positive rate (TPR), and positive prediction value (PPV). 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 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 5-fold cross validation was used for evaluating the prediction accuracy.

Analyte Annotation and Marker Identification

Table S2 presents 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 $[M+H]^+$, $[M+Na]^+$, $[M+K]^+$, $[M\text{-}H_2O+H]^+$, $[M+2Na\text{-}H]^+$, $[M+2K\text{-}H]^+$, and $[M+NH_4]^+$. 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. The average mass spectra of OSCC and HC are shown in **Fig. 1A** . From the average mass spectra, it can be clearly observed that the peaks at m/z 135.0029 (lactic acid, $[M+2Na\text{-}H]^+$), 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-stochastic neighbor embedding (t-SNE), was introduced to reduce the high-dimensional metabolite ions information into a three-dimensional (3D) feature space. In the constructed 3D feature space, serum cases from the same group were well clustered whereas those cases from different groups can be separated (**Fig. 1B**). This result demonstrated there exists a substantial difference in serum metabolic profiles that can be used for OSCC and HC prediction.

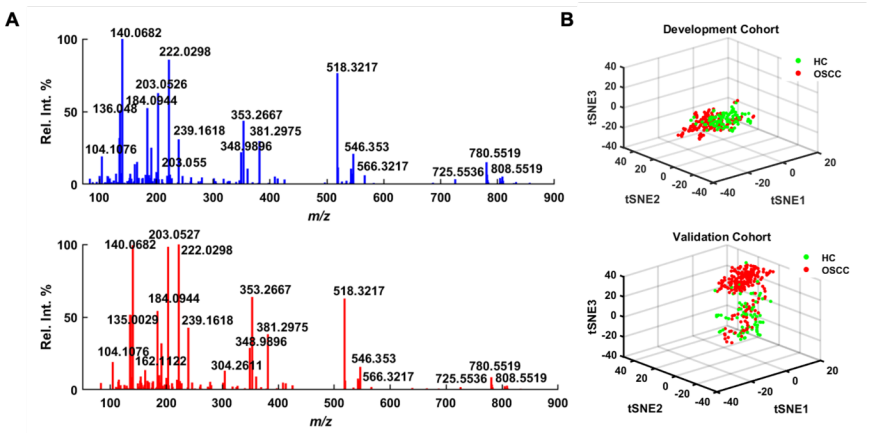


Figure 1. Serum metabolic profiling of OSCC and HC by t-SNE clustering: (A) The mass spectrum averaged from the HC (blue) and OSCC (red) groups, respectively. (B) metabolic profiling of first batch of 254 serum samples as the training dataset; and 565 serum samples as the test dataset.

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 (**Fig. 2A**). After removing ions that either were 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 S3**). A volcano graph highlighted the top 10 metabolites with the most obvious fold changes (**Fig. 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** displays the relative expressions of these top 10 metabolites in the form of a heatmap.

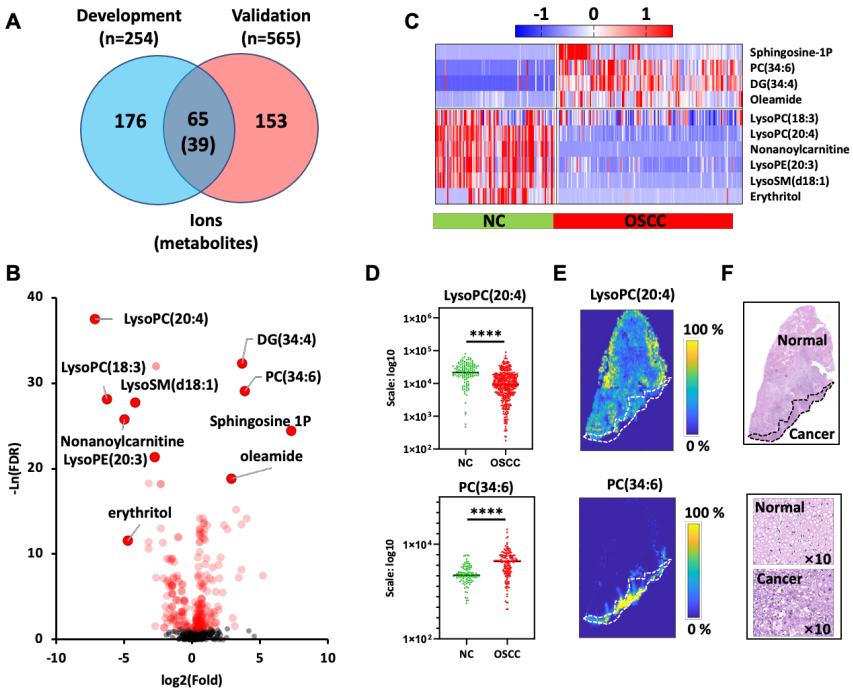


Figure 2. Discovery and validation of the serum metabolite markers. (A) Venn graph displayed the number of discriminating metabolite ions selected in the development and validation cohort and the number of common ions; (B) The most significantly metabolite ions with the largest fold changes were highlighted 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.

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 intact OSCC tumor tissues were prepared for DESI-MSI analysis. Then the spatial distribution of target metabolites across the resected OSCC tissues were mapped and compared. The *in situ* validation by DESI-MSI revealed that 8 out of top 10 metabolites showed the same trends of change 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 (**Figs. 2D, 2E,** and **2F**).

Expression of Salivary Metabolite Markers in Serum

Given the 106 characteristic metabolites previously studied in salivary metabolic profiling,[18] the extent of their changes in serum between 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 S4**), 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 with these 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 powers for OSCC screening 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 S5**), their performances on 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 and specificity both at 84 % (**Fig. 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 that 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.[33] Briefly, the absolute weights of the 66 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, 2 metabolite ions were composed and trained in the first cohort. As is shown in **Fig. 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 **Fig. 3C** and **Table S6** . 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 (**Fig. 3D**). The confusion matrix showed that the optimal SVM model can obtain a true positive rate at 94 % for OSCC detection (**Fig. 3E**). The final prediction accuracy reached 89.6 % on the test set (**Table S7**).

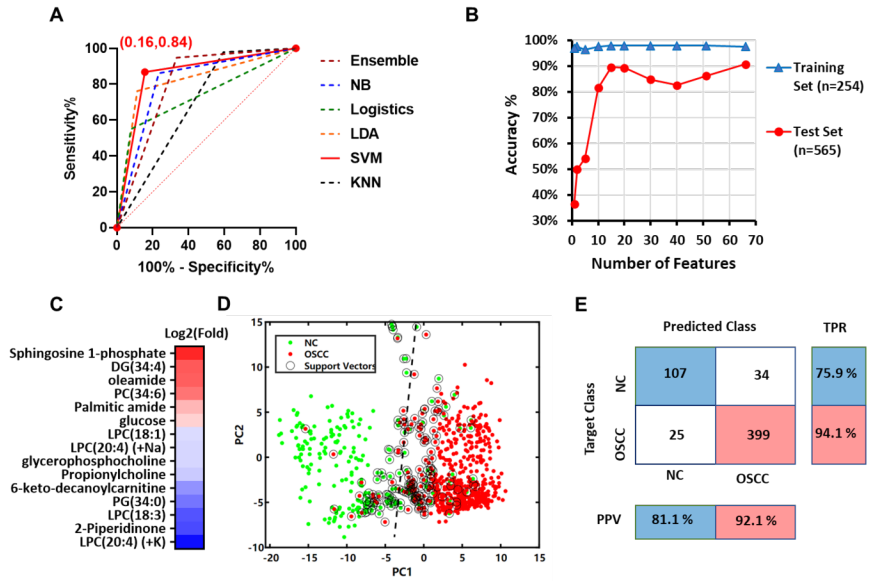


Figure 3. The development of serum metabolomics-based machine learning model for OSCC diagnosis. (A) Different machine learning models were initially investigated by comparing their diagnostic performance on the test set. SVM was chosen as the optimal one; (B) The number of features was investigated by 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 SVM were displayed in the feature space constructed with the first two principal components; (E) The classification result of the test set was displayed in a confusion matrix. Here TPR is the true positive rate and PPV is the positive prediction value.

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 in the development cohort (**Fig. 4A**). It can be seen the OSCC at T1 and T2 stages can be ideally separated from these samples at the T3 and T4 stages. 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 S8**). 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 4 out of the top 10 metabolites discovered in the univariate analysis (DG(34:2), oleamide, LPC(20:4), and nonanoyl carnitine) also contributed to this T1/T2 and T3/T4 stages discrimination (**Fig. 4B**). As shown in **Fig. 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).

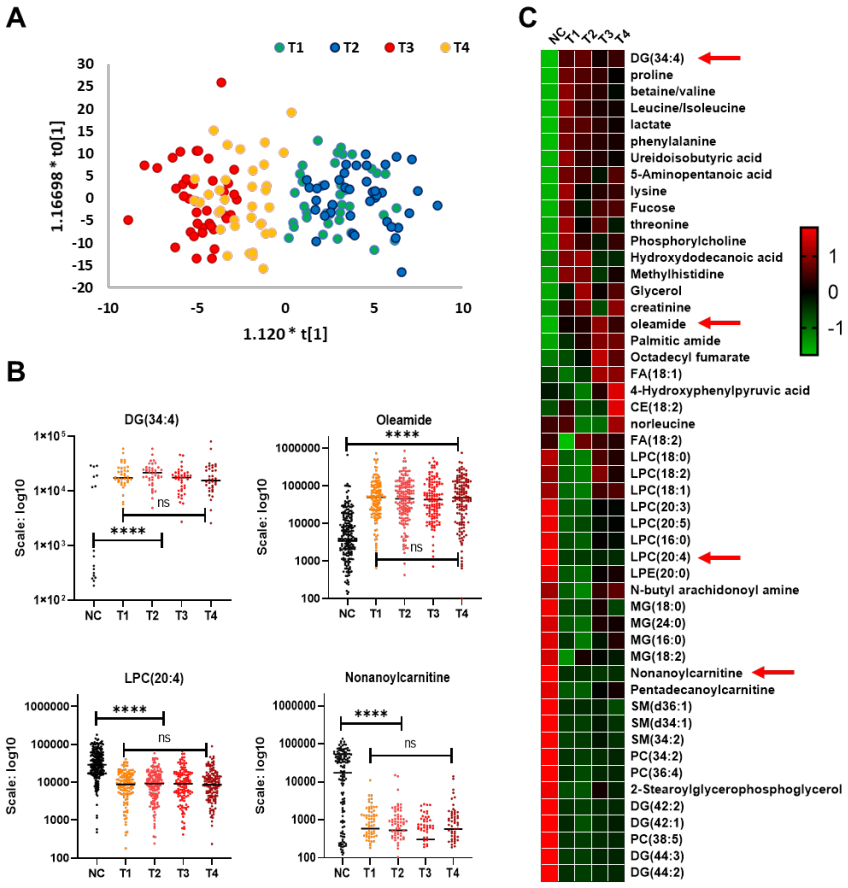


Figure 4. The OSCC staging by OPLS-DA according to the serum metabolic profiles. (A) The score plots of OSCC cases at different stages from T1 to T4. (B) The box plots for presenting the four lipid molecules' relative abundance across HC and OSCC at different stages. (C) The average metabolite profiles of HC and T1-T4 OSCC subgroups. The metabolites were selected by their VIP scores.

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 hours, 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.[34] 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 CSPI-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.

From the studies of serum metabolomics reported here and the previous saliva metabolomics, the OSCC-associated discriminating metabolites were identified, respectively. The pathway enrichment analysis revealed which metabolism pathways are influenced in serum and saliva (**Table S9**). The four representative metabolism 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 pathways including glycerolipid (GL), glycerophospholipid (GPL), and sphingomyelin (SM) (**Fig. 5**). According to the fold changes of these metabolites (**Tables S3** and **S4**), the changes of many metabolites become less obvious in serum, although the 57 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 metabolism 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**).

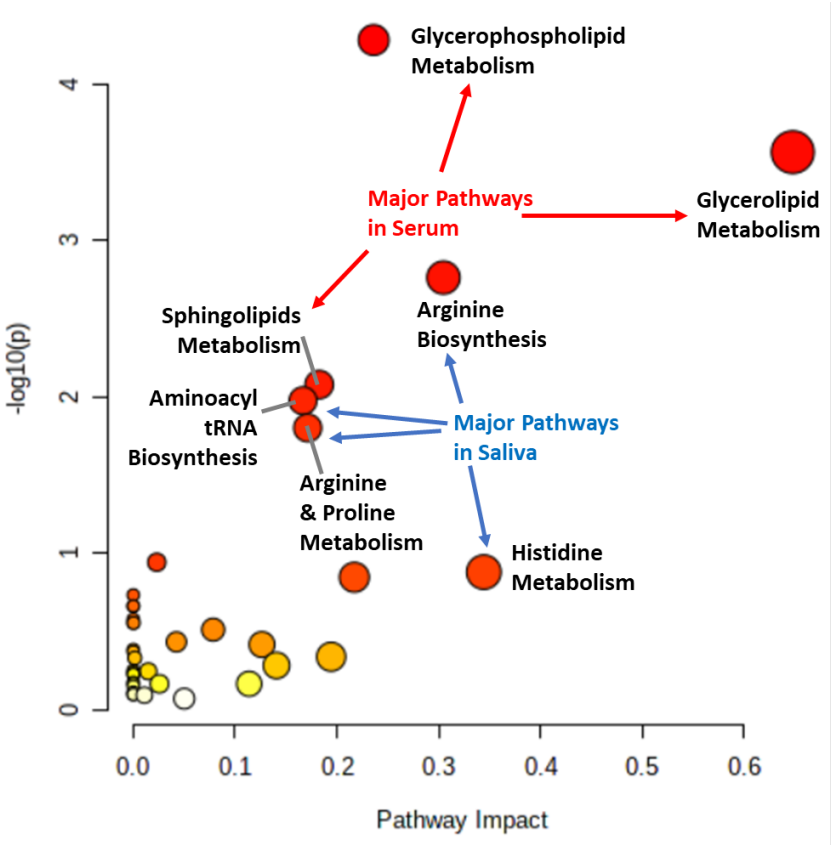


Figure 5. The OSCC-associated metabolism pathways and the involved major metabolites.

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.[35] Therefore, this inter-specimen derived difference in dysregulated metabolism 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.[36] 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.[37-39] The dysregulation in aminoacyl tRNA biosynthesis pathway hints at enhancing protein synthesis.[40, 41] 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), MG(16:0)], and acyl carnitine (e.g., acetyl carnitine, nonanoyl carnitine, 6-keto-decanoylcarnitine, 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.[42] The abnormal SPL metabolism (e.g., sphingosine 1-phosphate, sphinganine, and phytosphingosine) suggests cell proliferation dysregulation.[43] These dysregulated metabolite markers could not only serve as potential markers for OSCC diagnosis but also might assist in 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 analyte enrichment process.

It was also found that the serum metabolome has the potential for roughly assessing OSCC stages (**Fig. 4A**). Although so far, no inter-stage statistical significance was found among any single serum metabolite (**Fig. 4B**), a clear pattern difference appears especially when the OSCC stage was developed from T1 or T2 into T3 or T4 (**Fig. 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. The criteria of traditional hypothetical tests in univariate analysis (conventionally referred to P or FDR < 0.05) might be too cautious 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 founded 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 molecular diagnosis (89.6 % accuracy). All these findings indicate that CPSI-MS/ML can be a very useful tool for providing a simple, fast, affordable diagnostic method for OSCC screening.

Abbreviations

ANOVA: one-way analysis of variance; AUC: area under curve; cfDNA: cell-free deoxyribonucleic acid; CID: collision-induced dissociation; CPSI-MS: conductive polymer spray ionization mass spectrometry; CTC: cir-

culating tumor cell; DESI-MS: desorption electrospray ionization mass spectrometry; DG: diacylglycerol; DT: decision tree; FDR: false discovery rate; GPC: glycerophosphocholine; HC: healthy contrast; IS: internal standard; KNN: k-nearest neighboring; LDA: linear discriminant analysis; lyso-GPC: lyso-glycerophosphocholine; ML: machine learning; MS/MS: tandem mass spectrometry; mRNA: messenger ribonucleic acid; NB: naïve Bayesian; OPLS-DA: orthogonal partial least squares data analysis; OSCC: oral squamous cell carcinoma; PML: premalignant lesion; PPV: positive prediction value; QC: quality control; ROC: receiver operating characteristic; RSD: relative standard deviation; SVM: supporting vector machine; TIC: total ion current; t-SNE: t-stochastic neighbor embedding; TPR: true positive rate; VIP: variable importance on projection;

Acknowledgements

This work was funded by the National Natural Science Foundation (21974027); Jiangsu Province's Key Provincial Youth Talents Program (Grant No. QNRC2016841); Nanjing Municipal Key Medical Laboratory Constructional Project Funding (Since 2012); Nanjing Medical Science and Technique Development Foundation (YKK18123); and the US Air Force Office of Scientific Research through Basic Research Initiative (AFOSR FA9550-12-1-0400).

Data Availability Statement

Serum metabolomic profile data of all study cases are available upon requesting.

Ethical Statement

Human serums 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.

References

1. Globocan, International Agency for Research on Cancer, World Health Organization. Cancer Today 2020.

2. Marur S, D'Souza G, Westra WH, Forastiere AA. HPV-associated head and neck cancer: a virus-related cancer epidemic. *Lancet* 2010;11: 781-9.

3. Galvao-Moreira LV, da Cruz, MCFN. Saliva protein biomarkers and oral squamous cell carcinoma. *Proc Nat Acad Sci* 2017;114: E109-E10.

4. Hartwell LH. Reply to Galvao-Moreira and da Cruz: Saliva biomarkers to complement the visualization-based oral cancer detection. *Proc Nat Acad Sci*2017;114: E111.

5. Yakob M, Fuentes L, Wang MB, Abemayor E, Wong DTW. Salivary biomarkers for detection of oral squamous cell carcinoma - current state and recent advances. *Curr Oral Health Rep* 2014;1: 133-41.

6. Ge S, Zhou H, Zhou Z, Liu L, Lou J. Serum metabolite profiling of a 4-Nitroquinoline-1-oxide-induced experimental oral carcinogenesis model using gas chromatography-mass spectrometry. *PeerJ* 2021;9: e10619.

7. Xu H, Yang Y, Zhao H, Yang X, Luo Y, Ren Y, Liu W, Li N. Serum miR-483-5p: a novel diagnostic and prognostic biomarker for patients with oral squamous cell carcinoma. *Tumor Biol* 2016;37: 447–53.

8. Sun L, Liu L, Fu H, Wang Q, Shi Y. Association of decreased expression of serum mir-9 with poor prognosis of oral squamous cell carcinoma patients. *Med Sci Monit* 2016;22: 289-94.

9. Liu C, Lin J, Cheng H, Hsu Y, Cheng C, Lin S. Plasma miR-187* is a potential biomarker for oral carcinoma. *Clinical oral investigations* 2017;21: 1131–8.

10. Lu Y, Chen Y, Wang H, Tsai C, Chen W, Huang Y, Fan K, Tsai C, Huang S, Kang C, Chang JTC, Cheng AJ. Oncogenic function and early detection potential of miRNA-10b in oral cancer as identified by microRNA profilin. *Cancer Prev Res* 2012.

11. Liu C, Tsai M, Tu H, Lui M, Cheng W, Lin S. miR-196a overexpression and mir-196a2 gene polymorphism are prognostic predictors of oral carcinomas. *Ann Surg Oncol*2013;20: S406–S14.

12. Bu J, Bu X, Liu B, Chen F, Chen P. Increased expression of tissue/salivary transgelin mrna predicts poor prognosis in patients with oral squamous cell carcinoma (OSCC) surgery.*Med Sci Monit* 2015;21: 2275-81.

13. Aggarwal S, Sharma SC, Das SN. Galectin-1 and galectin-3: plausible tumourmarkers for oral squamous cell carcinoma and suitable targets for screening high-risk population.*Clinica Chimica Acta* 2015;442: 13-21.

14. Martinez-Outschoorn UE, Peiris-Pages M, Pestell RG, Sotgia F, Lisanti MP. Cancer metabolism: a therapeutic perspective. *Nat Rev Clinica Oncol* 2017;14: 11-31.

15. Zhou Z, Alvarez D, Milla C, Zare RN. Proof of concept for identifying cystic fibrosis from perspiration samples. *Proc Nat Acad Sci* 2019;116: 24408–12.

16. Pu F, Chiang S, Zhang W, Ouyang Z. Direct sampling mass spectrometry for clinical analysis. *The Analyst* 2019;144: 1034-51.

17. Yao YN, Di D, Yuan ZC, Wu L, Hu B. Schirmer Paper Noninvasive Microsampling for Direct Mass Spectrometry Analysis of Human Tears. *Ana Chem* 2020;92: 6207-12.

18. Mendes TPP, Pereira I, de Lima LAS, Morais CLM, Neves ACON, Martin FL, Lima KMG, Vaz BG. Paper Spray Ionization Mass Spectrometry as a Potential Tool for Early Diagnosis of Cervical Cancer. *J Am Soc Mass Spectrom* 2020;31: 1665-72.

19. Huang YC, Chung HH, Dutkiewicz EP, Chen CL, Hsieh HY, Chen BR, Wang MY, Hsu CC. Predicting Breast Cancer by Paper Spray Ion Mobility Spectrometry Mass Spectrometry and Machine Learning. *Anal Chem* 2020;92: 1653-7.

20. Vijayalakshmi K, Shankar V, Bain RM, Nolley R, Sonn GA, Kao CS, Zhao H, Tibshirani R, Zare RN, Brooks JD. Identification of diagnostic metabolic signatures in clear cell renal cell carcinoma using mass spectrometry imaging. *Int J Cancer*2020;147: 256-65.

21. Song X, Yang X, Narayanan R, Shankar V, Ethiraj S, Wang X, Duan N, Ni Y, Hu Q, Zare RN. Oral squamous cell carcinoma diagnosed from saliva metabolic profiling. *Proc Nat Acad Sci* 2020;17: 16167-73.

22. Song, X, Chen, H., Zare, R.N. Conductive polymer spray ionization mass spectrometry for biofluid analysis. *Anal Chem* 2018;90: 12878-85.

23. Rai V, Mukherjee R, Routray A, Ghosh AK, Roy S, Ghosh BP, Mandal PB, Bose S, Chakraborty C. Serum-based diagnostic prediction of oral submucous fibrosis using FTIR spectrometry. *Spectrochim Acta A: Mol Biomol Spectrosc* 2018 189: 322–9.

24. Saraswat M, Makitie A, Tohmola T, Dickinson A, Saraswat S, Joenvaara S, Renkonen S. Tongue cancer patients can be distinguished from healthy controls by specific n-glycopeptides found in serum. *Proteom Clin Appl* 2018;16: 1800061.

25. Yonezawa K, Nishiumii S, Kitamoto-Matsuda J, Fujita T, Morimoto K, Yamashita D, Saito M, Otsuki N, Irino Y, Shinohara M, Yoshida M, Nibu KI. Serum and tissue metabolomics of head and neck cancer. *Cancer Genom Proteom*2013;10: 233-8.

26. Kong X, Yang X, Zhou J, Chen S, Li X, Jian F, Deng P, Li W. Analysis of plasma metabolic biomarkers in the development of 4-nitroquinoline-1-oxide-induced oral carcinogenesis in rats. *Ontology Letters* 2015;9: 283-9.

27. Bag S, Banerjee DR, Basak A, Das AK, Pal M, Banerjee R, Paul RR, Chatterjee J. NMR (1H and 13C) based signatures of abnormal choline metabolism in oral squamous cell carcinoma with no prominent Warburg effect. *Biochemical and biophysical research communications* 2015;459: 574-8.

28. Tiziani S, Lopes V, Günther UL. Early stage diagnosis of oral cancer using 1H NMR–based metabolomics.*Neoplasia* 2009;11: 269–76.

29. Yang X, Zhang X, Jing Y, Ding L, Fu Y, Wang S, Hu S, Zhang L, Huang X, Ni Y, Hu Q. Amino acids signatures of distance-related surgical margins of oral squamous cell carcinoma.*EBioMedicine* 2019;48: 81-91.

30. Yang X, Jing Y, Wang S, Deng F, Zhang X, Chen S, Zheng L, Hu Q, Ni Y. Integrated non-targeted and targeted metabolomics uncovers amino acid markers of oral squamous cell carcinoma. *Front Pharmacol* 2020;10.

31. Hrydziuszko O, Viant MR. Missing values in mass spectrometry based metabolomics:

an undervalued step in the data processing pipeline. Metabolomics. 2012; 8: S161–S174.

32. Benjamini, Y., Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. Royal Stat. Soc. 1995; 57:289–300.

33. Kohavi R, John GH. Wrappers for feature subset selection. *Artif Intell* 1997;97: 273-324.

34. Jennifer A. Kirwan, RJMW, David I. Broadhurst, Mark R. Viant. Direct infusion mass spectrometry metabolomics dataset: a benchmark for data processing and quality control. *Sci Data* 2014;1: 140012.

35. Vander Heiden MG, DeBerardinis RJ. Understanding the intersections between metabolism and cancer biology. *Cell* 2017;168: 657-69.

36. Pavlova NN, Thompson CB. The emerging hallmarks of cancer metabolism. *Cell Metab* 2016;23: 27-47.

37. Zhu J, Thompson CB. Metabolic regulation of cell growth and proliferation. *Nat Rev Mol Cell Biol* 2019;20: 436-50.

38. Ward PS, Thompson CB. Metabolic reprogramming: a cancer hallmark even warburg did not anticipate.*Cancer cell* 2012;21: 297-308.

39. Liu JY, Wellen KE. Advances into understanding metabolites as signaling molecules in cancer progression.*Current opinion in cell biology* 2020;63: 144-53.

40. Park SG, Schimmel P, Kim S. Aminoacyl tRNA synthetases and their connections to disease. *Proc Nat Acad Sci* 2008;105: 11043-9.

41. Kim S, You S, Hwang D. Aminoacyl-tRNA synthetases and tumorigenesis: more than housekeeping.*Nat Rev Cancer* 2011;11: 708-18.

42. Corbet C, Feron O. Emerging roles of lipid metabolism in cancer progression. *Current opinion in clinical nutrition and metabolic care* 2017;20: 254-60.

43. Ogretmen B. Sphingolipid metabolism in cancer signalling and therapy. *Nat Rev Cancer*2018;18: 33-50.