SUPPLEMENTARY INFORMATION

Oral Squamous Cell Carcinoma Diagnosed from Saliva Metabolic Profiling

Xiaowei Song[1*], Xihu Yang[2*], Rahul Narayanan[1], Vishnu Shankar[3], Sathiyaraj Ethiraj[1], Xiang Wang[2], Ning Duan[2], Yan-Hong Ni[4], Qingang Hu[2#], Richard N. Zare[1,3#]

1.Department of Chemistry, Fudan University, Shanghai, 200438, China
2.Department of Oral and Maxillofacial Surgery, Nanjing Stomatological Hospital, Medical School of Nanjing University, Nanjing, Jiangsu 210000, China
3. Department of Chemistry, Stanford University, Stanford, California 94305 USA
4. Central Laboratory of Stomatology, Nanjing Stomatological Hospital, Medical School of Nanjing University, Nanjing, Jiangsu 210000, China

* These authors contribute equally to this work.

#Correspondence

Richard N. Zare, rnz@stanford.edu; Qingang Hu, qghu@nju.edu.cn

Table of Contents

www.pnas.org/cgi/doi/10.1073/pnas.2001395117

Materials and Methods

Saliva Collection and Pretreatment. All of the 373 saliva samples were collected from the Department of Oral and Maxillofacial Surgery, Nanjing Stomatological Hospital. The clinical information on OSCC patients, PML patients, and HC volunteers were summarized in **Table S1**. These volunteers diagnosed to be complicated with other oral diseases (e.g., chronic periodontitis). Besides, saliva collected from another nine healthy volunteers were also treated as the negative quality control in pattern recognition of metabolic profiling and method validation. The saliva collection was approved by the medical ethics committee of the Nanjing Stomatology Hospital. All patients were informed and signed consent forms. To avoid diet interferences, mouth rinsing with ultrapure water was required before saliva collection. Oral hygiene products (e.g., toothpaste) were also not allowed for use before 1.0 h prior to sample collection. Whole saliva (500 µL) was harvested into an EP tube without exogenous stimulus. After centrifugation at 5000 rpm for 3 min, the supernatant was transferred and saved at -80 °C until use. Additionally, to confirm the discovered metabolites at the *in situ* level, tumor tissues collected from 22 OSCC patients were cryo-sectioned (15 μm) for DESI-MSI confirmation.

CPSI-MS and DESI-MSI method. For CPSI-MS analysis, the general procedure was according to that previously reported. Briefly, the 3 µL saliva (spiked with 3-choloro-phenylalanine as internal standard) was first micropipetted onto the conductive polymer tip, which was tuned by XYZ positioner and set at 8.0 mm distance away from the MS inlet. When the saliva was dried to form a spot, methanol-water (7:3, v/v, 3 µL) was used as the spraying solvent to dissolve endogenous metabolites in the dried spot. When the + 4.5 kV high voltage was applied onto the conductive polymer by a copper alligator clip, a plume of charged microdroplets will be sprayed and carry the components into the mass spectrometer. The LTQ Orbitrap Velos mass spectrometer (Thermo Scientific) was employed for the ambient MS analysis task. The full scan mode was used for untargeted metabolic profiling within the range of *m/z* 50-500 under positive mode. The MS capillary temperature was set at 275 °C with the S-lens voltage set at 55 V. The automatic gain control was set at 3E6 with the maximum injection time set at 400 ms.

For tissue imaging, a commercial 2D DESI system (Prosolia, Inc, US) was employed in the positive ion mode with all of the other MS parameters same as above. Acetonitrile-water (7:3, v/v) was used as the spray solvent with the flow rate set at 2.0 µL/min under nebulizer gas pressure of 1.0 MPa. The impact angle between sprayer head (+4.0 kV applied) and substrate was 55°. The height of sprayer tip and the distance from tip to transport tube were all set at 4.5 mm.

CPSI-MS and DESI-MSI Data Preprocessing. The Xcalibur software was employed for generating average mass spectra and converting a batch of raw data files into cdf files. The ion's *m/z* within ±0.005 Da mass tolerance will be defined with one mass bin. Only the mass bin that was successfully detected among more than 30 % of samples was used for data matrix construction. Self-written MATLAB 2019a (Mathworks, US) script was used to automatically extract average peak intensities of

MS scans in the sample's time window, constructing the data matrix which consisted of m samples (rows) and n metabolite ions (columns). To eliminate the influence of signal response fluctuation on the statistical analysis, both total ion current (TIC) and internal standard ion intensity (*m/z* 222.03, $[M+Na]^+$) of each sample's average mass spectrum can be the optional choice for intensity normalization to achieve the good modeling performance. The matrix was transferred through natural logarithm and then standardized to be centered at zero with standard deviation scaled at one, ruling out the magnitude's biasing influence on the classification. As for DESI-MSI data, Massimager (Chemmind Technologies Co., Ltd, China) and a self-programmed MATLAB script was used for ion image reconstruction and spatial segmentation.

Machine Learning. The samples were divided into two batches for CPSI-MS data acquisition separately within two different periods. The collected data was split into 193 samples in the first batch for training and 180 samples in the second batch for validation. The first batch contained saliva from 65 healthy contrast volunteers (HC), 64 patients with premalignant lesions (PML), and 64 oral squamous cell carcinoma patients (OSCC). The second batch contained saliva from 60 HC, 60 PML, and 60 OSCC cases. We first trained the model via cross-validation (20-folds) on the training set and externally validate the model performance on the held-out 5% test set.

A total of 627 common peaks were extracted from two batches (training and validation datasets) of saliva mass spectra. The MATLAB2019a was employed for developing machine learning models to differentiate the HC, PML, and OSCC cases. The in-built "classification learner" and "regression learner" APPs were employed for investigate and compare the model performances in fitting and generalization. The investigated classification models included decision tree (DT), discriminant analysis (DA), support vector machine (SVM), K nearest neighbor (KNN), naïve Bayesian classifier (NB). Besides, "neural net pattern recognition" APP was also used to build the artificial neural network classification model. The in-built "Lasso" function in MATLAB was used to establish the Lasso regression model. The number of cross-validation was set at 20 folds using the in-built "crossvalind" function. The data matrix of the training and validation datasets as well as the related self-written code and functions have been uploaded into the open-source platform OSF. All of these files can be accessed from the following linkage: https://osf.io/nv32d/.

The accuracy and mean squared error (MSE) were used to evaluate the performance of the different machine learning models. Receiver operating curve (ROC) analysis was carried out using Graphpad Prism to evaluate the diagnostic metrics including area under curve (AUC), specificity, and sensitivity. Confusion matrix was used to display the classification results for the training and validation datasets.

Dynamic Simulation. After finishing the machine learning training and validation, the Lasso model was deployed under the Simulink platform to simulate the on-line automated screening of the HC, PML, and OSCC population. After the conversion of acquired batch raw files into cdf format, the feature ions intensities of each MS scan were automated extracted, transformed, and input into the deployed Lasso model to give the instant diagnosis result. The in-built functional blocks such as "from workspace," "sumover," "matrix multiply," "sum," "constant," "scope," and "switch" in Simulink platform were organized to stimulate the dynamic Lasso recognition model. More details about the configuration are shown in **Figure S10**.

Statistical Analysis. Univariate analysis was first implemented to search for significantly changed metabolite ions among HC, PML, and OSCC groups using student t test. The P values were checked and adjusted with the false discovery rate (FDR) using Benjamini-Hochberg method. The ions were picked out if the fold change was over 2.0 or less than 0.5 ($P < 0.05$ and FDR<0.1). For pattern recognition of different groups, SIMCA-P (Umetrics, Umea, Sweden) was used for (orthogonal) partial least squares discriminant analysis ((O)PLS-DA) of metabolic profiles. Variables with importance in projection (VIP) larger than 1.5 were considered to make a high contribution to the classification. In addition, OPLS-DA were also used to investigate the inter-time, inter-day, and individual variation of the metabolic profile, as well as the influence of diet on sample classification.

Metabolite Identification and Metabolic Pathway Searching. The significantly changed ions in HR-MS were first identified through database searching from HMDB (http://hmdb.ca/). To achieve the elemental composition and possible list of endogenous metabolites, the relative error of exact *m/z* value was limited to 5.0 ppm. The type of ion adducts were limited to $[M+H]^+$, $[M+Na]^+$, $[M+K]^+$, $[M-H_2O+H]^+$, $[M+2Na-H]^+$, $[M+2K-H]^+$, $[M+NH4]^+$ under positive mode. The fragment ions produced under CID-MS2 were also used to assign the exact structure for a specific metabolite. The CID-MS2 fragmentation patterns were compared either with the standards, or matched with the self-built CID spectra collections, or the standard MS2 spectra in the metabolomics database (HMDB). The identified metabolites of interest were put into the open-source platform, MetaboAnalyst (https://www.metaboanalyst.ca), to search for these altered metabolic pathways.

Comparison to Flow Injection-ESI and Paper Spray Ionization-MS. Practically, flow injection (FI)-ESI consumes more biological fluid samples (20~500 μL) or it dilutes the biofluid for filling up the syringe. Direct injection of biofluid will seriously foul the tubing system and cause inter-samples cross talk unless the syringe and tubing system are thoroughly washed after each assay. This limits its practical use in large scale sample tests. The saliva will strongly suppress the generation of electrospray at the ESI capillary outlet due to its strong viscosity and surface tension. In comparison, CPSI-MS utilizes self-conductive materials, and only consumes no more than several μL of sample on the tip. The materials are cheap and can be either disposable or repeatedly used after simple wiping with wet, dust-free tissue. Thus, CPSI-MS is very suitable for large scale metabolomics screening. Compared to paper spray ionization (PSI), we have discussed this before at length(1). The signal intensity for polar or hydrophilic species tends to be at least 20 to 100 fold higher in CPSI-MS than those in the PSI-MS. We have added the above information to the supporting information.

Reference

1. Song XW, Chen, H., Zare, R.N. (2018) Conductive Polymer Spray Ionization Mass Spectrometry for Biofluid Analysis. *Anal Chem* 90:12878-12885.

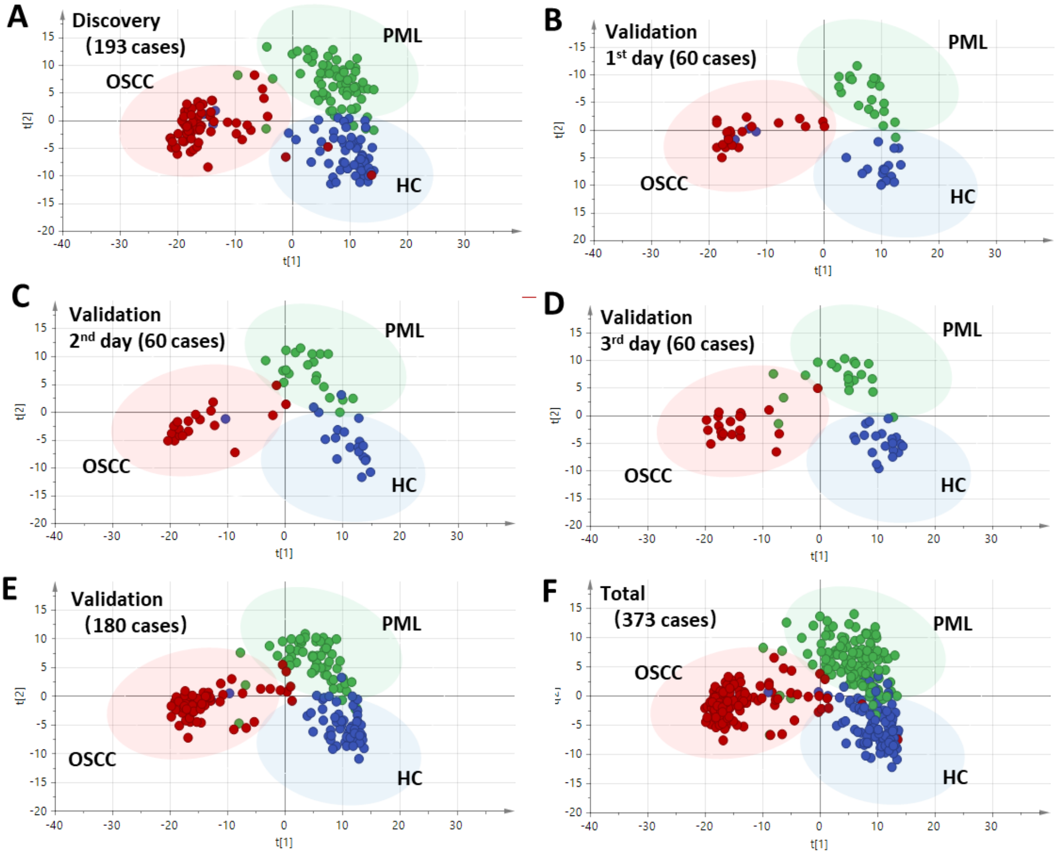
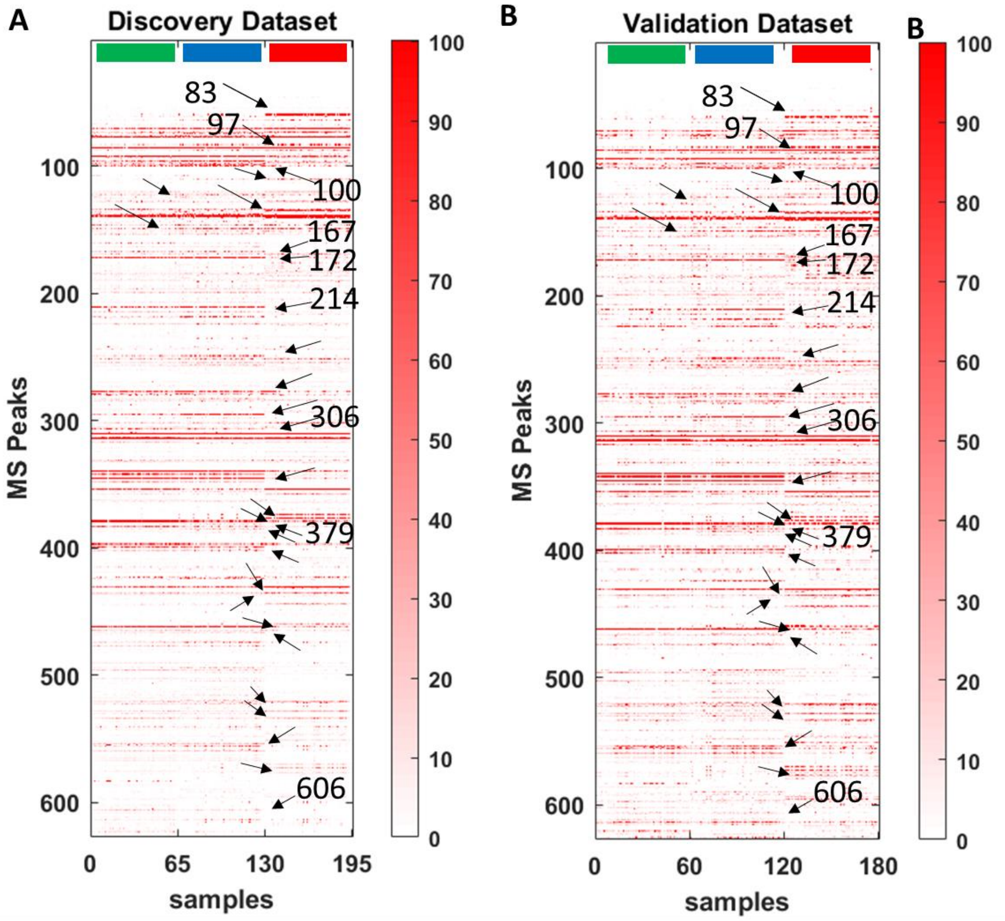


Figure S1. Salivary metabolic profiling of different batches and groups with OPLS-DA: (A) metabolic profiling of first batch of 193 saliva cases as the discovery dataset; (B-D) metabolic profiling of three sub-batches of saliva cases over successive three days; (E) saliva metabolic profiling of 180 validation cases; and (F) metabolic profiling of total 373 saliva cases.



C

Pearson Coefficient	HC_ discovery	PML_ discovery	OSCC_ discovery	HC_ validation	PML_ validation	OSCC_ validation
HC_ Discovery	1.00					
PML_ Discovery	0.81	1.00				
OSCC_ Discovery	0.66	0.66	1.00			
HC_ Validation	**0.95**	0.74	0.51	1.00		
PML_ Validation	0.77	**0.90**	0.44	0.81	1.00	
OSCC_ Validation	0.74	0.63	**0.86**	0.67	0.50	1.00

Figure S2. Visual display of the mass spectra of all saliva cases acquired with CPSI-MS to evaluate the repeatability of CPSI-MS results. (A) First batch of 193 cases as the discovery dataset. (B) Second batch of 180 cases as the validation dataset. The obviously changed MS peaks are indicated by arrows. The annotated peaks were ones which can be identified as cadaverine (peak No.83), glycerol (No.97), 5-aminopentanoate (No. 100), proline (No. 167), betaine (No. 172), arginine (No.

214), glucose (No. 306), phosphorylcholine (No.379), and MG(18:0/0:0/0:0) (No.606). (C) Pearson correlation coefficients.

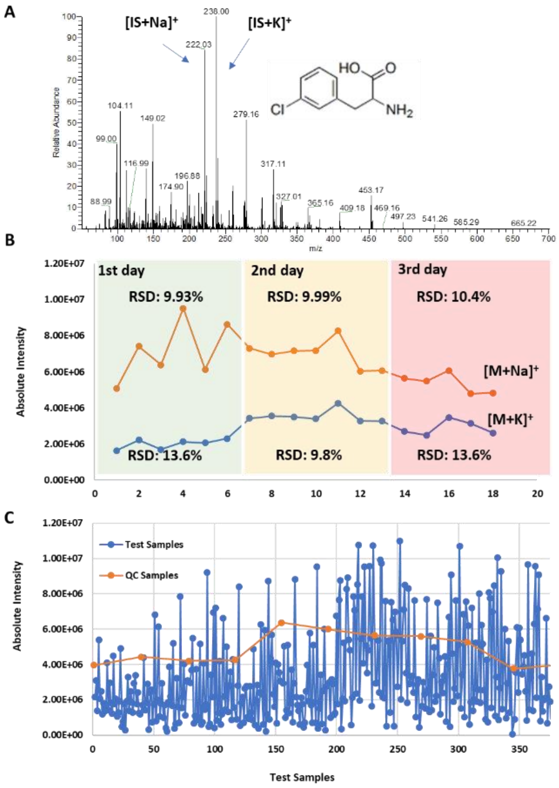


Figure S3. Monitoring and quality control of the salivary metabolic data acquisition with CPSI-MS. (A) Representative mass spectrum of saliva spiked with internal standard (IS) 3-cholrophenylalanine (IS peaks at $[M+Na]^+$ *m/z* 222.03 and $[M+K]^+$ *m/z* 238.00). (B) The intra-day and inter-day variation of IS peak ions in 18 QC saliva samples. (C) Variation of internal standard peak ($[M+Na]^+$ *m/z* 222.0383) spiked in the tested and quality control saliva samples. The IS ion intensities fall into the range of mean±SD (3.34E6±2.5E6). The QC samples came from the unique IS-spiked pooling aliquot collected from 20 normal contrast saliva.

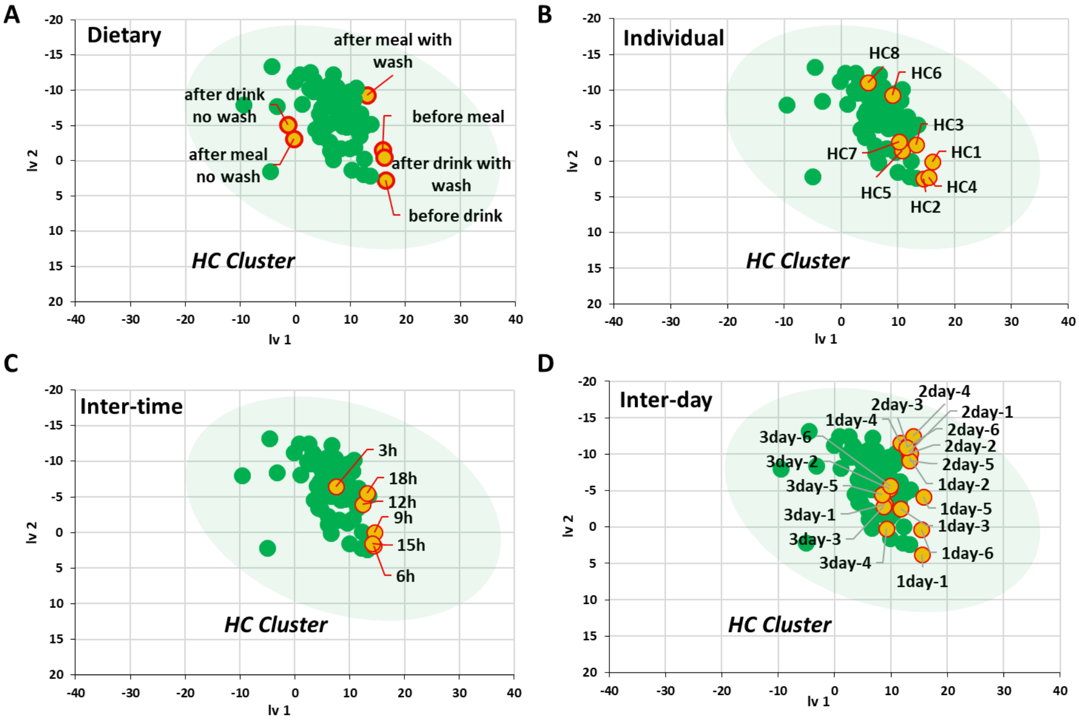


Figure S4. Investigation of CPSI-MS/OPLS-DA robustness to the (A) dietary, (B) individual, (C) inter-time, and (D) inter-day variation in salivary components. HC represents healthy contrast saliva collected from eight different persons. The collection was set at fixed date and time. Mouth rinsing with ultrapure water was required before saliva collection to avoid diet interferences. Oral hygiene products (e.g., toothpaste) were also not allowed for at least 1.0 h prior to sample collection.

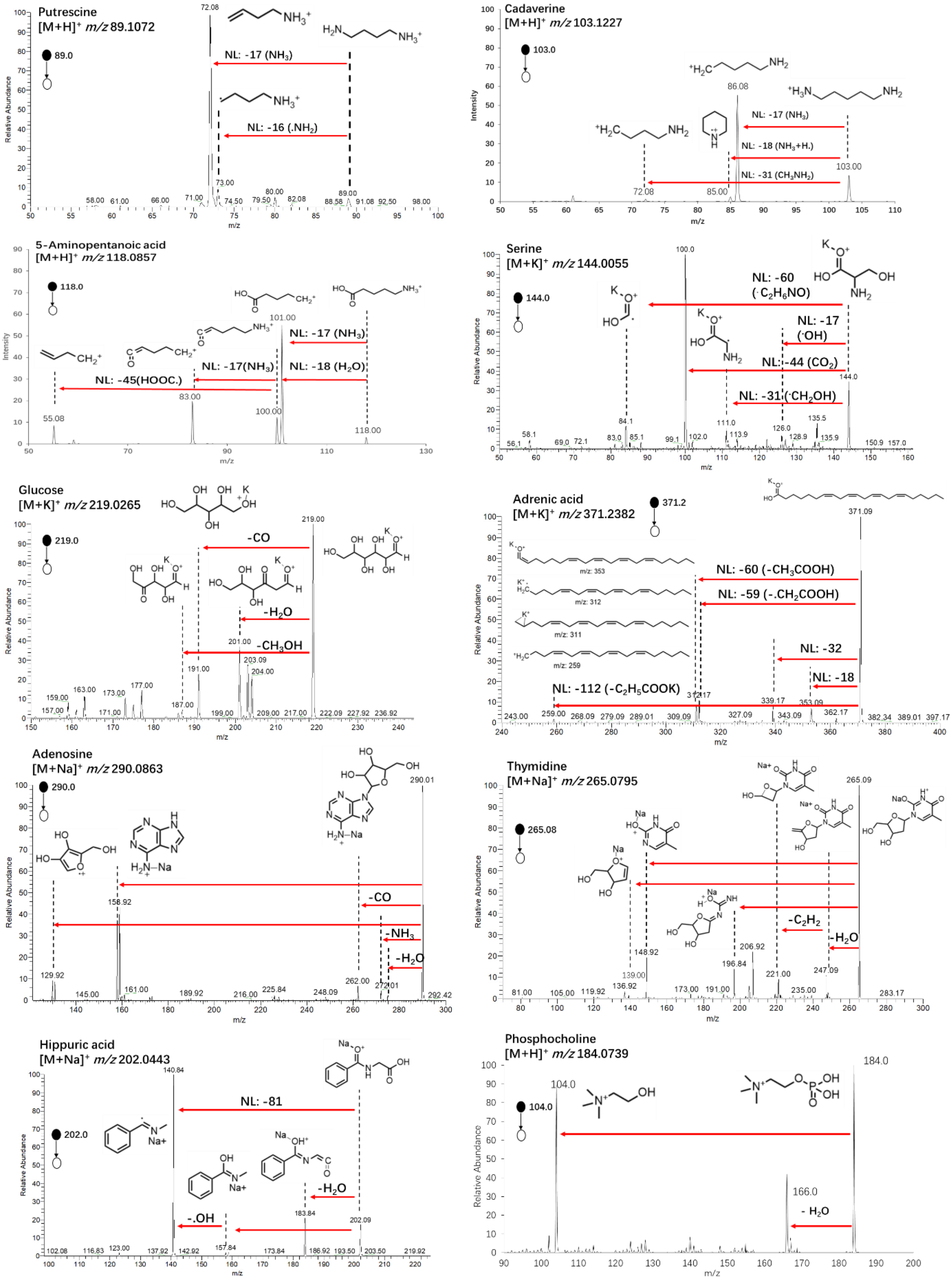


Figure S5. Top 10 metabolites identified according to exact fragment ion assignments given by CID-MS/MS.

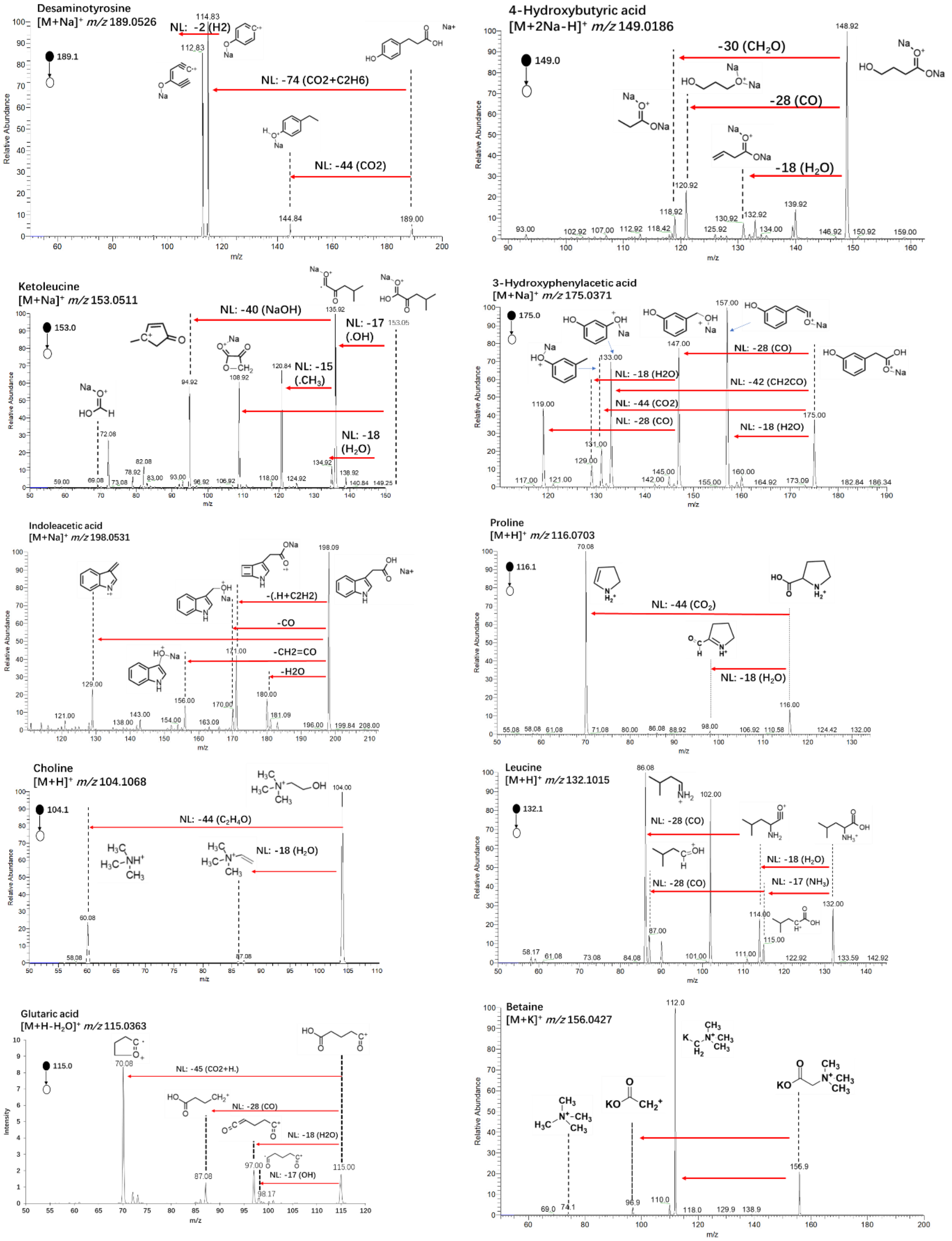


Figure S6. Representative metabolites identified according to exact fragment ion assignments given by CID-MS/MS.

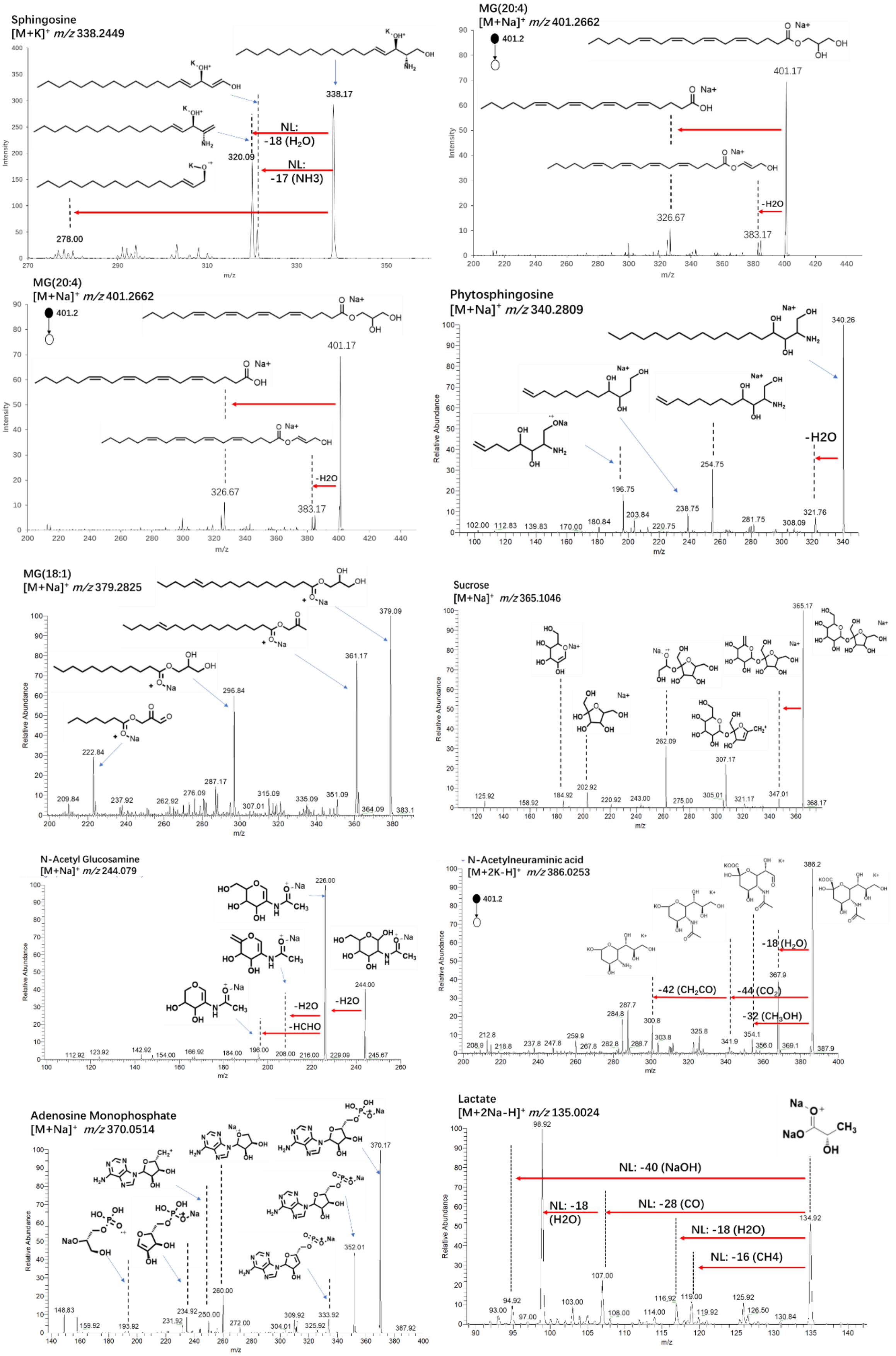


Figure S6 (continued).

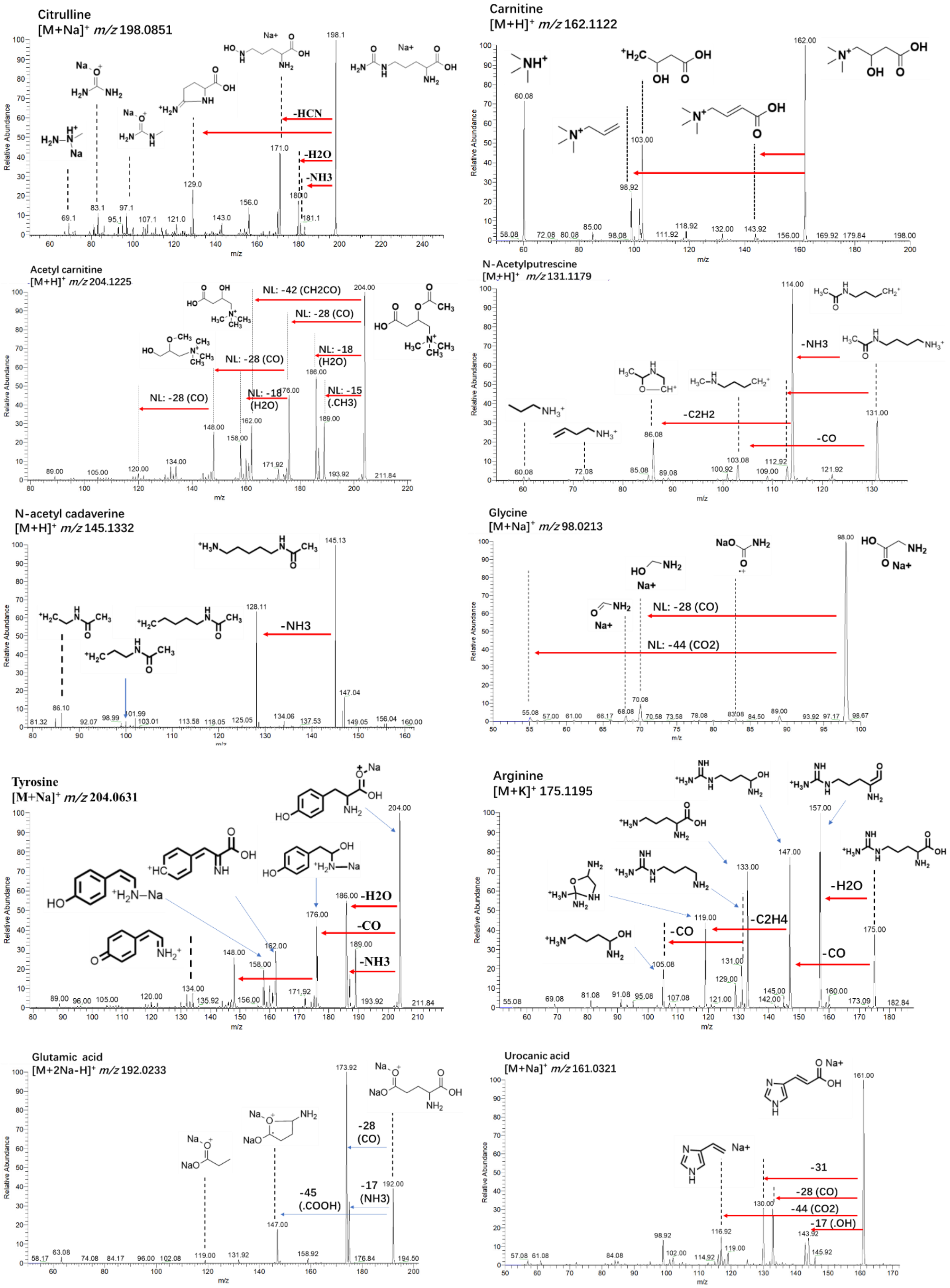


Figure S6 (continued).

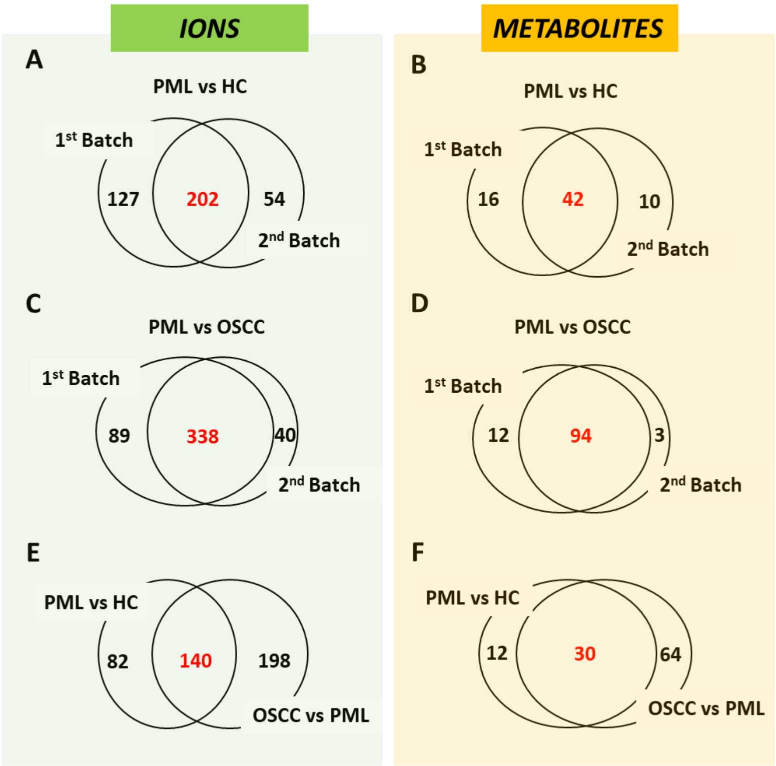


Figure S7. Venn diagrams of mutual ions or metabolites among different batches or groups. The mutual ions (A) and metabolites (B) with significant changes during premalignant progression that were discovered by the first and second batches. The mutual ions (C) and metabolites (D) with significant changes during malignant progression that were discovered by the first and second batches. (E) The mutual ions which were found significantly changed both in premalignant and malignant stages. (F) The mutual metabolites which were found significantly changed both in premalignant and malignant stages.

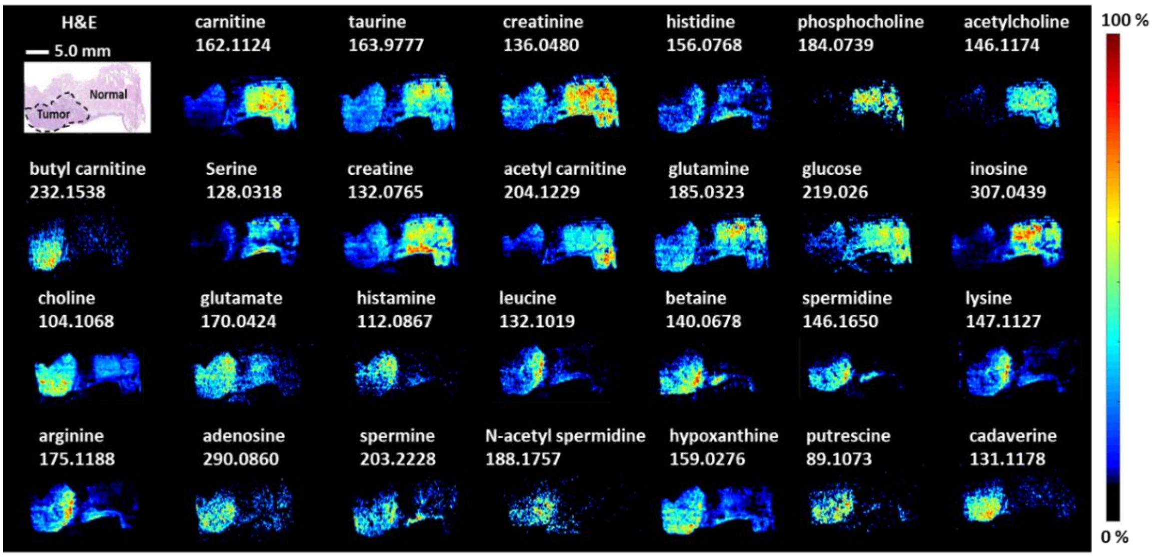


Figure S8. Confirmation of the discovered metabolites in saliva at the primary carcinoma site by DESI-MSI.

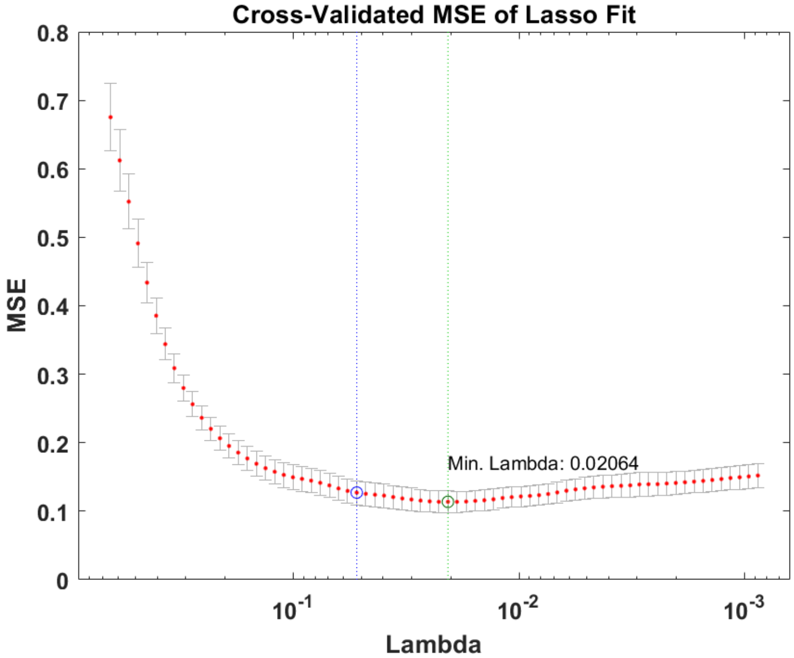


Figure S9. The MSE changes with the lambda during the 11th round of Lasso model training.

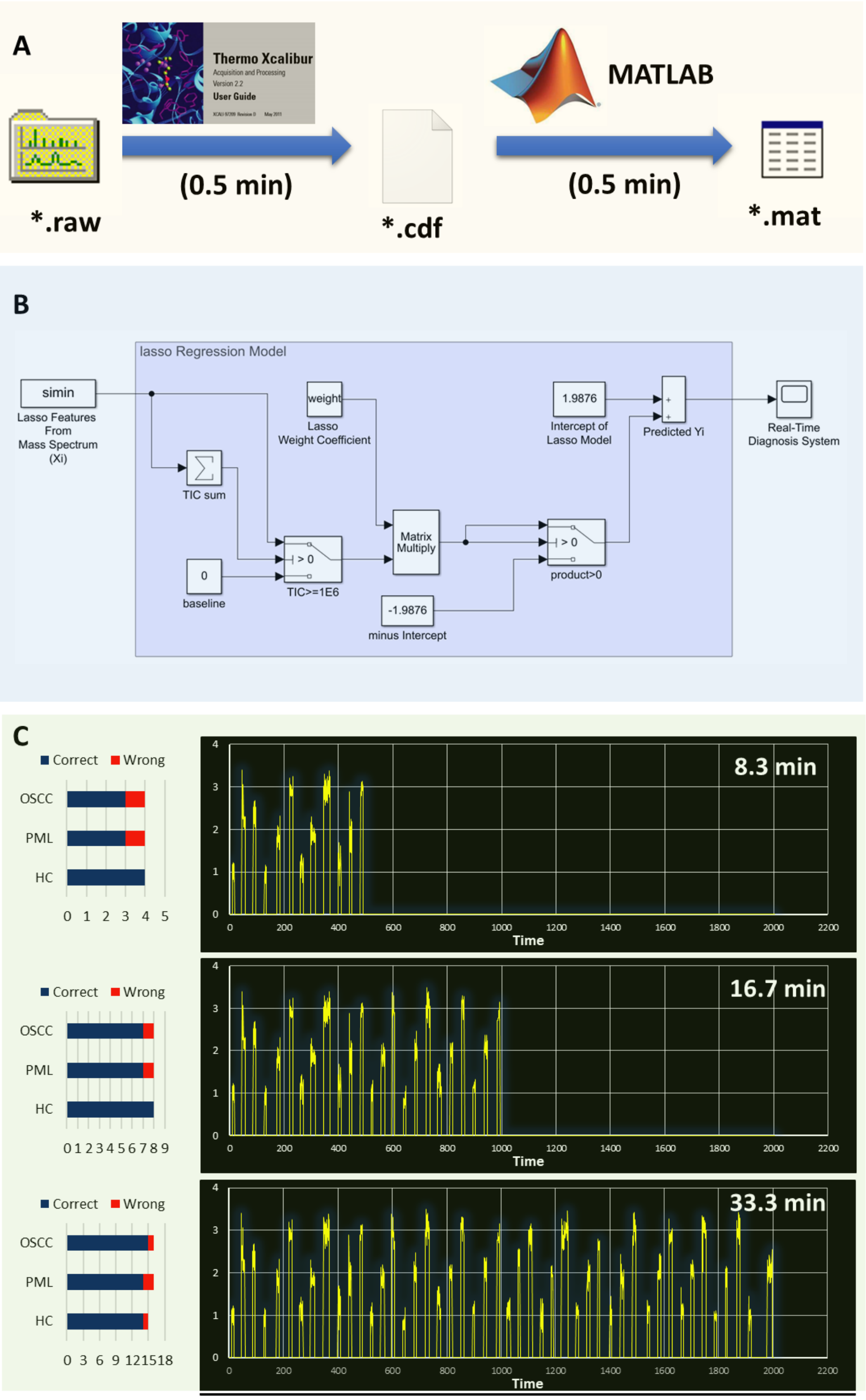


Figure S10. Pipeline of nearly real-time molecular diagnosis of OSCC by CPSI-MS/ML using the MATLAB/Simulink system. (A) The pipeline for data format conversion from .raw to .cdf followed by importing into MATLAB for automatic data processing including metabolic feature extraction and prediction. (B) The Lasso regression model consisted of different function blocks in Simulink to stimulate the scan-by-scan molecular diagnosis at nearly real-time. (C) The simulated real-time diagnosis process at three different times.

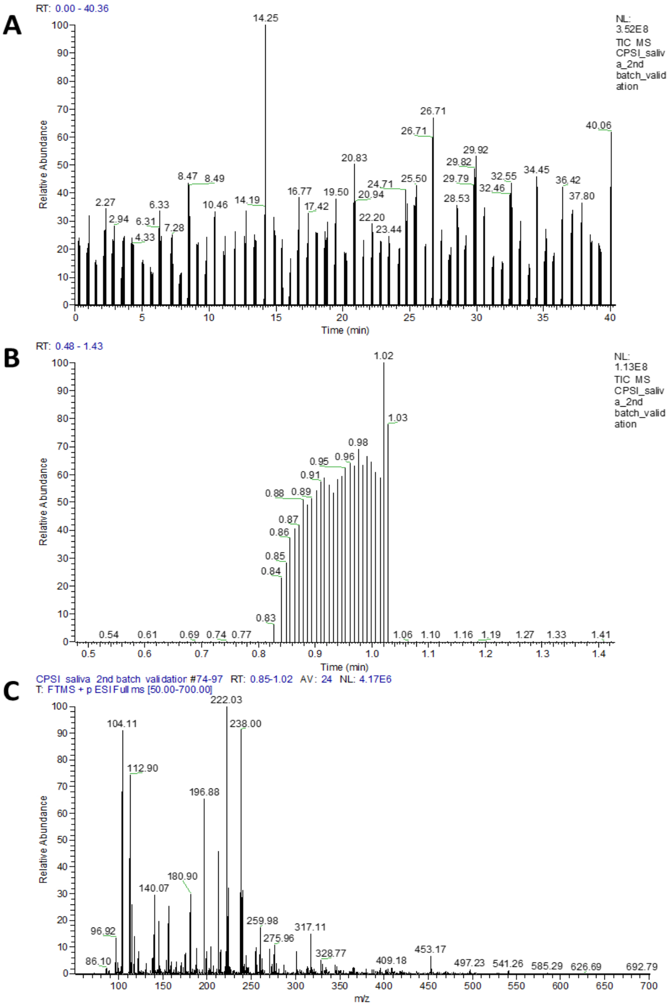


Figure S11. The representative high-throughput MS collection from the validation batch. (A) Representative TIC graph for high throughput screening of 60 samples within 40 minutes; (B) The zoomed TIC focused on single case with the time window of 0.83-1.03 minutes; (C) The mass spectrum averaged from the scans with that time window.

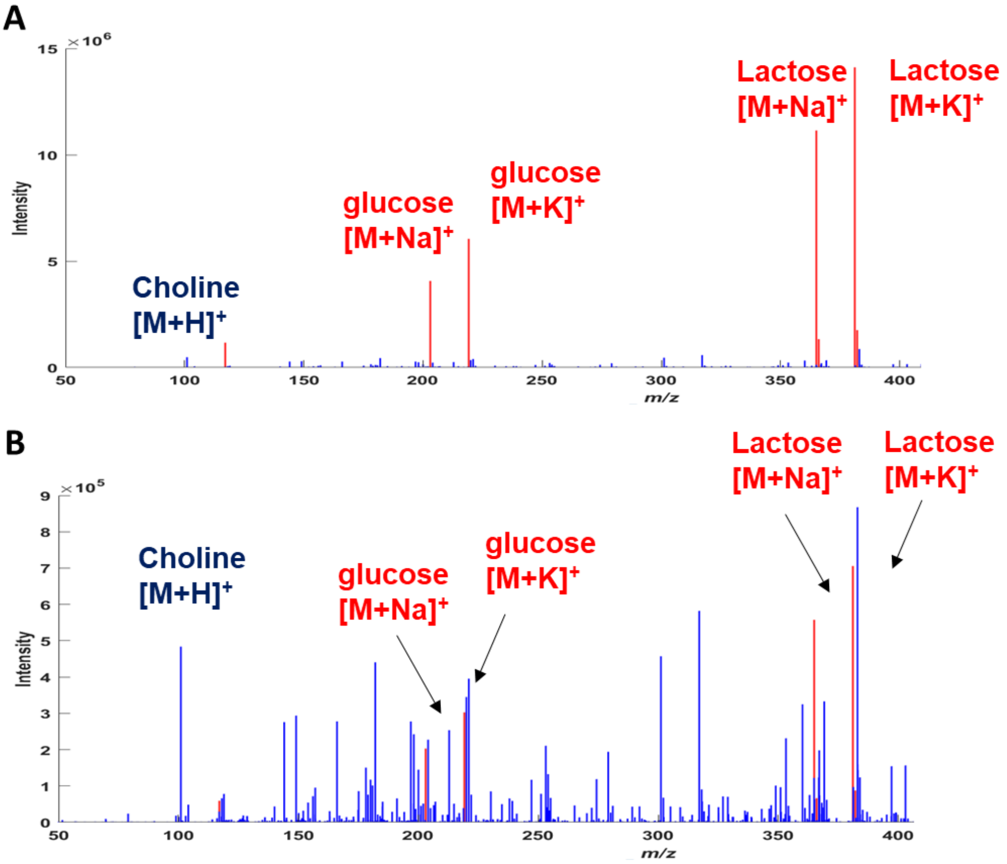


Figure S12. Mass spectra of saliva collected after meal directly (A) and after mouth pre-rinsing (B).

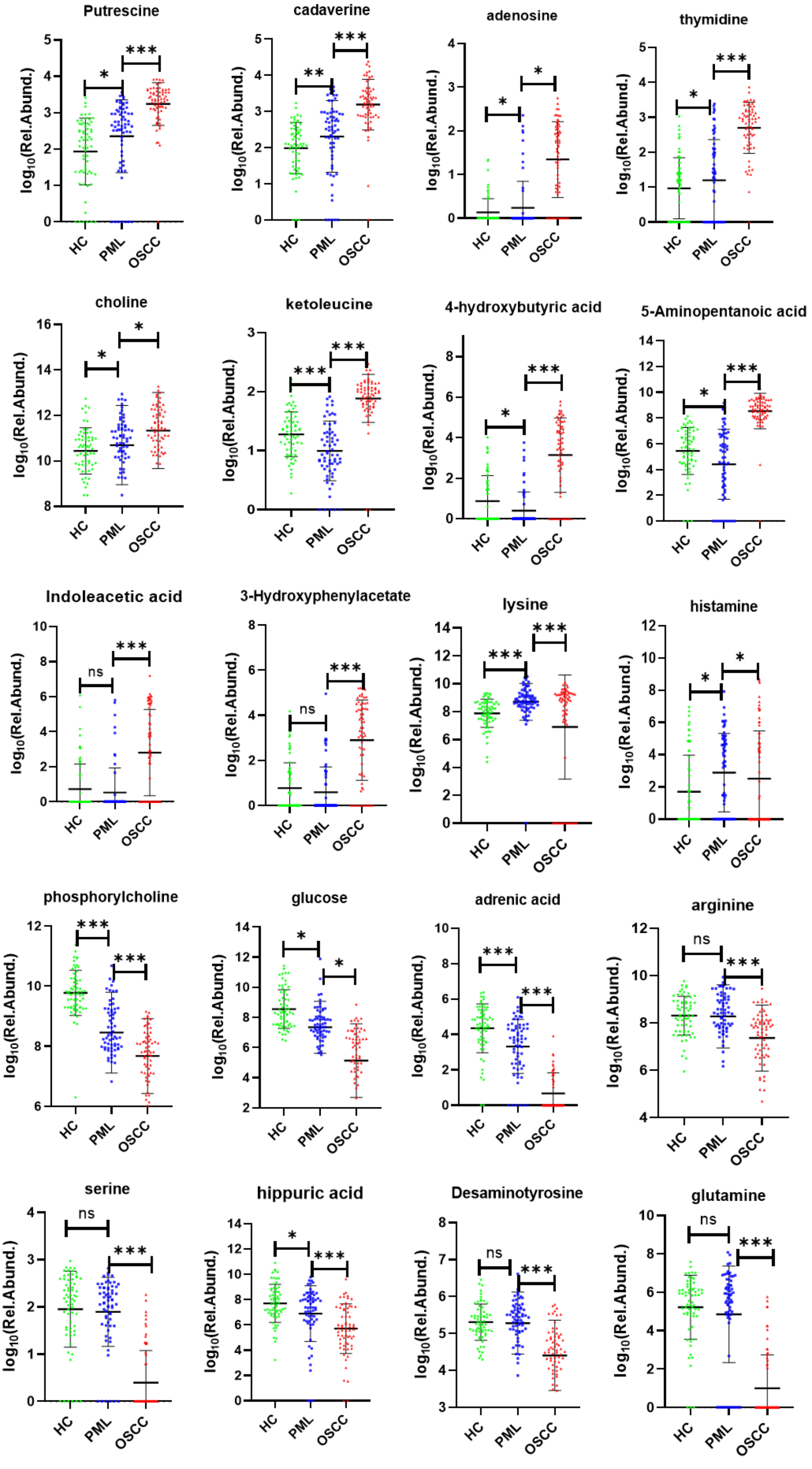
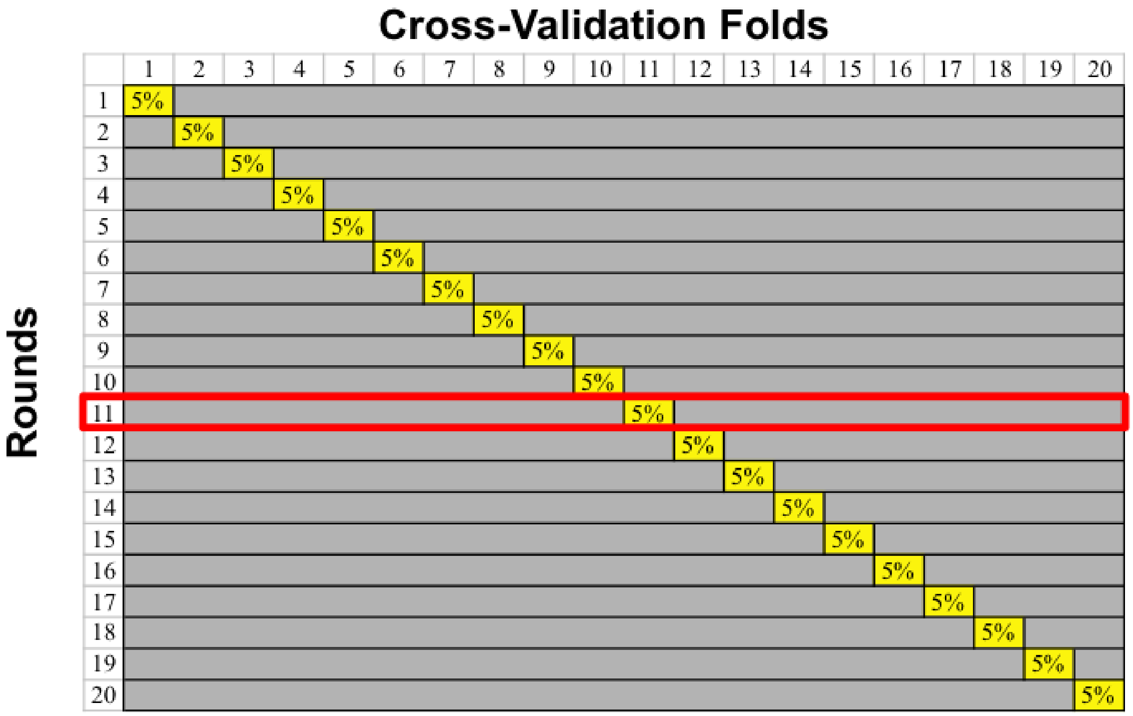


Figure S13. Box plots of representative metabolites that had continuous change tendencies from HC to OSCC.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	5%																			
2		5%																		
3			5%																	
4				5%																
5					5%															
6						5%														
7							5%													
8								5%												
9									5%											
10										5%										
11											5%									
12												5%								
13													5%							
14														5%						
15															5%					
16																5%				
17																	5%			
18																		5%		
19																			5%	
20																				5%

Figure S14. Schematic of cross-validation procedure. The yellow 5% represents the test fold in each round of cross-validation, while the rest in gray corresponds to the training fold. The selected model from round 11 has been indicated by the red box.

Table S1. Summary of information on OSCC patients, PML patients, and HC volunteers.

Sample Batch	Characteristics	OSCC	PML	HC
Discovery	Race	Chinese	Chinese	Chinese
	Numbers	65	64	64
	Age (Range)	35-65	35-65	30-60
	Gender (M/F)	35/30	34/30	34/30
	Prior-therapy	N	N	----
	Stages	Stage: I: 15 II: 21 III: 12 IV: 17	Subtype OLP: 40 OLK: 24	----
	Tumor sites	Tongue (21), cheek (12), jaw (3), mouth floor (6), gums (9), palate (4), lips (2), oropharynx (7)		
Validation	Race	Chinese	Chinese	Chinese
	Numbers	60	60	60
	Age (Range)	35-65	35-65	30-60
	Gender (M/F)	30/30	30/30	30/30
	Prior-therapy	N	N	----
	Stages:	Stage I: 14 II: 19 III: 11 IV: 16	Subtype: OLP: 40 OLK: 20	----
	Tumor sites	Tongue (19), jaw (3), cheek (12), mouth floor (6), gums (9), palate (4),lip (1), oropharynx (6),		

Table S2. Tentative assignment of metabolite ions.

Metabolite	Formula	Adduct	Theo. m/z	Exp. m/z	Delta (ppm)
1,3-dimethyluracil	C6H8N2O2	$[M+K]^+$	179.0217	179.0211	-3.35
1-methylhistidine	C7H11N3O2	$[M+2K-H]^+$	246.0042	246.0046	1.63
2-hydroxyvaleric acid	C5H10O3	$[M+H-H_2O]^+$	101.0603	101.06	-2.97
2-ketobutyric acid	C4H6O3	$[M+Na]^+$	125.0209	125.0203	-4.80
3-hydroxyphenylacetate	C8H8O3	$[M+2Na-H]^+$	197.0185	197.0189	2.03
4-aminobutyrate	C4H9NO2	$[M+Na]^+$	126.052	126.0514	-4.74
4-hydroxybutyric acid	C4H8O3	$[M+2Na-H]^+$	149.0185	149.0186	0.67
5-aminopentanoic acid	C5H11NO2	$[M+2Na-H]^+$	162.0501	162.0494	-4.32
8-hydroxy-7-Methylguanine	C6H7N5O2	$[M+2Na-H]^+$	226.0311	226.0317	2.65
8-oxoguanine	C5H3N5O2	$[M+NH_4]^+$	183.0625	183.062	-2.71
Acetyl carnitine	C9H17NO4	$[M+H]^+$	204.123	204.1225	-2.45
Acetyl carnosine	C11H16N4O4	$[M+H]^+$	269.1244	269.1249	1.86
Acetylcholine	C7H16NO2	$[M+H]^+$	146.1176	146.1171	-3.12
Adenosine	C10H13N5O4	$[M+Na]^+$	290.086	290.0863	1.03
AMP	C10H14N5O7P	$[M+Na]^+$	370.0523	370.0514	-2.43
Adipic acid	C6H10O4	$[M+H-H_2O]^+$	129.0552	129.0547	-3.87
Adrenic acid	C22H36O2	$[M+K]^+$	371.2347	371.2355	2.15
Allantoin	C4H6N4O3	$[M+Na]^+$	181.0331	181.0327	-2.44
Arginine	C6H14N4O2	$[M+K]^+$	175.119	175.1195	2.86
Aspartate	C4H7NO4	$[M+2Na-H]^+$	178.0086	178.0081	-2.99
Betaine	C5H11NO2	$[M+Na]^+$	140.0682	140.0678	-2.86
Cadaverine	C5H14N2	$[M+H]^+$	103.1230	103.1227	-2.91
Caprylic acid	C8H16O2	$[M+2K-H]^+$	221.0341	221.0338	-1.36
Carnitine	C7H15NO3	$[M+H]^+$	162.1122	162.1123	0.62
Choline	C5H14NO	$[M+H]^+$	104.1072	104.1068	-3.84
Citrulline	C6H13N3O3	$[M+Na]^+$	198.0849	198.0851	0.96
Creatine	C4H9N3O2	$[M+Na]^+$	154.0587	154.0583	-2.60
Creatinine	C4H7N3O	$[M+Na]^+$	136.0480	136.0475	-3.68
Cytosine	C4H5N3O	$[M+2K-H]^+$	187.9623	187.9615	-4.26
Decenoic acid	C10H18O2	$[M+Na]^+$	193.1199	193.1200	0.52
Deoxycholic acid	C24H40O4	$[M+H]^+$	393.2999	393.2991	-2.03
Desaminotyrosine	C9H10O3	$[M+Na]^+$	189.0522	189.0526	2.12
Dihydrothymine	C5H8N2O2	$[M+Na]^+$	151.0473	151.0474	0.86
Dimethylarginine	C8H18N4O2	$[M+H]^+$	203.1503	203.1498	-2.46
Eicosatrienoic acid	C20H34O2	$[M+K]^+$	345.219	345.2178	-3.48
Ethanolamine	C2H7NO	$[M+H]^+$	62.0601	62.06	-0.97
Glucose	C6H12O6	$[M+2Na-H]^+$	225.0345	225.0336	-4.00

Table S2 (Continued)

Metabolite	Formula	Adduct	Theo. m/z	Exp. m/z	Delta (ppm)
Glutamate	C5H9NO4	$[M+2Na-H]^+$	192.0239	192.0233	-3.33
Glutamine	C5H10N2O3	$[M+Na]^+$	169.0584	169.0582	-1.18
Glutaric acid	C5H8O4	$[M+H-H_2O]^+$	115.0396	115.0393	-2.61
Glycerol	C6H14O5	$[M+Na]^+$	115.0366	115.0361	-4.02
glycerol-3-phosphate	C3H9O6P	$[M+H]^+$	173.0210	173.0207	-1.58
Glycerophosphocholine	C8H20NO6P	$[M+K]^+$	296.0660	296.0646	-4.73
Glycine	C2H5NO2	$[M+Na]^+$	98.02125	98.0212	-0.51
Guanosine	C10H13N5O5	$[M+Na]^+$	306.0809	306.0804	-1.63
Hippuric acid	C9H9NO3	$[M+Na]^+$	202.0475	202.0469	-2.97
Histidine	C6H9N3O2	$[M+Na]^+$	178.0587	178.0581	-3.15
Histamine	C5H9N3	$[M+H]^+$	112.0869	112.0865	-3.57
Hydroxyarachidonic acid	C20H32O3	$[M+2Na-H]^+$	365.2063	365.2061	-0.55
Hydroxydodecanedioic acid	C12H22O5	$[M+2Na-H]^+$	291.1179	291.1187	2.75
Hydroxydodecanoic acid	C12H24O3	$[M+Na]^+$	239.1618	239.1607	-4.60
Hydroxyoctanoic acid	C8H16O3	$[M+K]^+$	199.0731	199.0723	-4.02
Hypoxanthine	C5H4N4O	$[M+H]^+$	137.0458	137.0458	0.00
Indoleacetic acid	C10H9NO2	$[M+2Na-H]^+$	220.0345	220.0335	-3.18
Inosine	C10H12N4O5	$[M+K]^+$	307.0439	307.0426	2.28
Ketoleucine	C6H10O3	$[M+Na]^+$	153.0522	153.0515	-4.57
Lactate	C3H5O3	$[M+2Na-H]^+$	135.0023	135.0017	-4.44
Leucic acid	C6H12O3	$[M+K]^+$	171.0418	171.0415	-1.77
Leucine	C6H13NO2	$[M+H]^+$	132.1019	132.1016	-2.23
Linoleic acid	C18H32O2	$[M+Na]^+$	303.2294	303.2288	-1.98
Linolenic acid	C18H30O2	$[M+K]^+$	317.1877	317.1865	-3.78
Lysine	C6H14N2O2	$[M+K]^+$	147.1128	147.1123	-3.43
Methionine	C5H11O2NS	$[M+Na]^+$	172.0403	172.04	-1.55
Methyladenine	C6H7N5	$[M+NH_4]^+$	167.1040	167.1037	-1.60
MG(14:0/0:0/0:0)	C17H34O4	$[M+Na]^+$	325.234	325.2332	-2.32
MG(16:0/0:0/0:0)	C19H38O4	$[M+Na]^+$	353.2662	353.2648	-3.96
MG(16:1/0:0/0:0)	C19H36O4	$[M+Na]^+$	351.2496	351.2492	-1.28
MG(18:0/0:0/0:0)	C21H42O4	$[M+K]^+$	397.2715	397.2704	-2.77
MG(18:1/0:0/0:0)	C21H40O4	$[M+Na]^+$	379.2819	379.2815	-1.05
MG(20:4/0:0/0:0)	C23H38O4	$[M+Na]^+$	401.2662	401.2657	-1.25
N1,N8-diacetyl spermidine	C11H23N3O2	$[M+H]^+$	230.1863	230.1855	-3.48
N1,N12-Diacetylspermine	C14H30N4O2	$[M+H]^+$	287.2442	287.2438	-1.39
N1-acetyl spermidine	C9H21N3O	$[M+H]^+$	188.1757	188.1751	-3.19

Table S2 (Continued)

Metabolite	**Formula**	**Adduct**	**Theo. m/z**	**Exp. m/z**	**Delta (ppm)**
N-acetyl neuraminic acid	C11H19NO9	[M+2K-H]+	386.025	386.024	-2.59
N-acetyl cadaverine	C7H16N2O	[M+H]$^+$	145.1335	145.1332	-2.07
N-acetyl putrescine	C6H14N2O	[M+H]$^+$	131.1179	131.1178	-0.76
N-acetyl glucosamine	C8H15NO6	[M+Na]$^+$	244.0792	244.0797	2.048
N-glycolylneuraminic acid	C11H19NO10	[M+H]$^+$	326.1082	326.1072	-3.06
N6,N6,N6-trimethyl-lysine	C9H20N2O2	[M+H-H_2O]$^+$	171.1498	171.1490	-4.67
Niacinamide	C6H6N2O	[M+H]$^+$	123.0554	123.0549	-3.78
Oleic acid	C18H34O2	[M+2Na-H]$^+$	327.2270	327.2258	-3.79
Ornithine	C5H12N2O2	[M+Na]+	155.0791	155.0785	-3.68
Palmitic acid	C16H32O2	[M+2Na-H]$^+$	301.2114	301.2104	-3.32
Palmitic amide	C16H33NO	[M+Na]$^+$	278.2454	278.2447	-2.52
Pentadecanoylcarnitine	C22H43NO4	[M+K]$^+$	424.2824	424.2815	-2.12
Phenylalanine	C9H11NO2	[M+Na]$^+$	188.0682	188.0676	-3.17
Phosphocreatine	C4H10N3O5P	[M+Na]$^+$	234.025	234.0244	-2.56
Phosphoethanolamine	C2H8NO4P	[M+2K-H]+	217.9381	217.9373	-3.67
Phosphorylcholine	C5H15NO4P	[M+2K-H]+	259.9856	259.9846	-3.85
Phosphoserine	C3H8NO6P	[M+Na]$^+$	207.9981	207.998	-0.48
Pipecolic acid	C6H11NO2	[M+H]$^+$	130.0863	130.0865	1.88
Piperidine	C5H11N	[M+H]$^+$	86.0964	86.0962	-2.32
Proline	C5H9NO2	[M+H]$^+$	116.0706	116.0703	-2.58
Propionyl carnitine	C10H19NO4	[M+H]$^+$	218.1387	218.1383	-1.83
Putrescine	C4H12N2	[M+H]$^+$	89.10732	89.1072	-1.35
Phytosphingosine	C18H39NO3	[M+H]$^+$	318.3003	318.2997	-1.89
Ribulose	C5H10O5	[M+Na]$^+$	173.0415	173.0412	-1.85
Ricinoleic acid	C18H34O3	[M+K]$^+$	337.214	337.2128	-3.56
Serine	C3H7NO3	[M+K]$^+$	144.0058	144.0055	-2.08
Spermidine	C7H19N3	[M+H]$^+$	146.1650	146.1648	-1.37
Sphinganine	C18H39NO2	[M+H]$^+$	302.3054	302.3046	-2.65
Sphingosine	C18H37NO2	[M+K]$^+$	338.2456	338.2449	-2.07
Sucrose	C12H22O11	[M+Na]$^+$	365.1054	365.1046	-2.19
Taurine	C2H7NO3S	[M+Na]$^+$	148.0035	148.0032	-2.03
Threonine	C4H9NO3	[M+Na]$^+$	142.0471	142.0465	-4.35
Thymidine	C10H14N2O5	[M+H]$^+$	243.0975	243.0983	3.29
Tryptophan	C11H12N2O2	[M+Na]$^+$	227.0791	227.0794	1.34
Tyrosine	C9H11NO3	[M+Na]$^+$	204.0631	204.0627	-1.96
Undecanoylcholine	C16H34NO2	[M+H]$^+$	272.259	272.2584	-2.20
Urea	CH4N2O	[M+K]$^+$	98.99552	98.9955	-0.20
Uridine	C9H12N2O6	[M+H]$^+$	245.078	245.077	-4.00
Urocanic acid	C6H6N2O2	[M+Na]$^+$	161.0321	161.0321	0.17

Bold annotations represent those metabolite that identified with CID-MS/MS experiment.

Table S3. Summary of significantly changed metabolites between healthy and premalignant lesions.

Metabolites	P value	FC*	Metabolites	P value	FC
Phosphoethanolamine	3.98E-06	9.62	Arginine	6.03E-07	2.19
Adenosine	4.21E-02	8.95	Acetylcholine	4.90E-05	2.19
N-acetyl cadaverine	0.000739	7.27	N1-acetyl spermidine	0.002755	2.19
Putrescine	0.001907	6.88	acetyl carnitine	0.017593	2.17
N-glycolylneuraminic acid	0.017109	6.61	Proline	0.015615	2.09
Piperidine	2.89E-05	4.92	linolenic acid	6.82E-03	0.48
Ethanolamine	0.002459	4.57	sphinganine	1.08E-13	0.48
Cadaverine	0.0001	4.1	hypoxanthine	0.002386	0.47
1,3-dimethyluracil	0.001041	3.48	1-methylhistidine	4.98E-03	0.46
Glutarate	0.007729	3.24	phosphoserine	7.36E-06	0.44
Niacinamide	0.002798	3.19	phosphorylcholine	4.82E-16	0.43
Propionyl carnitine	0.026664	3.15	aspartate	0.000613	0.35
N-Acetyl putrescine	0.007744	2.89	Lactate	6.95E-05	0.35
Phenylalanine	5.46E-09	2.67	palmitic acid	0.003698	0.24
Lysine	3.24E-05	2.62	MG(14:0/0:0/0:0)	0.000771	0.2
Thymidine	0.032123	2.55	linoleic acid	0.00097	0.19
N-acetylglucosamine	0.006244	2.42	sphingosine	3.90E-04	0.16
Choline	0.002355	2.35	Glucose	0.012314	0.14
N-acetylneuraminic acid	0.003376	2.28	8-hydroxy-7-methylguanine	6.89E-09	0.1
Histidine	0.000775	2.25	pentadecanoyl carnitine	1.98E-06	0.08
Allantoin	7.91E-05	2.24	Sucrose	0.045772	0

***FC represent fold change of PML versus HC, only metabolites with FC values larger than 2.0 or smaller than 0.5 were listed in the table.**

Table S4. Summary of significantly changed metabolites between premalignant lesions and oral squamous cell carcinoma.

Metabolites	P value	Fold Change	Metabolites	P value	Fold Change
4-hydroxybutyric acid	4.50E-12	31.07	phytosphingosine	0.013228	4.53
palmitic acid	2.20E-07	28.76	Carnitine	0.000167	2.31
propionyl carnitine	5.76E-11	18.80	N-acetylneuraminic acid	0.004	2.25
Guanosine	0.000303	16.41	MG(16:0/0:0/0:0)	2.46E-06	2.25
3-hydroxyphenylacetate	5.46E-11	15.86	methionine	0.004313	2.24
Adenosine	3.21E-07	12.96	Spermidine	0.010989	2.20
Serine	2.67E-06	11.01	8-hydroxy-7-methylguanine	0.016307	2.16
Lactate	2.75E-17	10.38	N-acetylglucosamine	0.005911	2.05
Phosphocreatine	0.000192	9.74	2-hydroxyvaleric acid	0.018276	0.48
pentadecanoyl carnitine	0.017002	9.69	hydroxyoctanoic acid	0.0027	0.48
Inosine	1.75E-06	8.80	Desaminotyrosine	2.66E-11	0.46
indoleacetic acid	0.000102	8.76	Phosphoserine	0.009535	0.42
MG(14:0/0:0/0:0)	6.92E-06	8.72	hippuric acid	0.000666	0.38
5-aminopentanoic acid	1.94E-14	6.60	leucic acid	6.08E-07	0.29
Leucine	3.18E-05	6.40	pipecolic acid	3.95E-11	0.29
Ketoleucine	1.33E-20	5.91	1,3-dimethyluracil	0.034503	0.28
1-methylhistidine	2.31E-15	5.78	Arginine	6.74E-10	0.27
oleic acid	1.33E-05	5.72	acetyl carnitine	0.00188	0.26
Cadaverine	0.003115	5.51	Creatine	3.84E-06	0.26
8-oxoguanine	1.72E-27	5.21	3-hydroxydodecanedioic acid	0.0029	0.25
deoxycholic acid	0.000639	5.09	Creatinine	1.13E-08	0.24
urocanic acid	1.60E-05	4.95	Tryptophan	2.03E-06	0.17
linoleic acid	3.08E-05	4.93	Butyrylcarnitine	0.010358	0.15
2-ketobutyric acid	1.30E-16	4.75	N-glycolylneuraminic acid	0.021629	0.14
decenoic acid	1.69E-20	4.71	adenosine monophosphate	4.81E-05	0.12
palmitic amide	0.002419	4.42	Glucose	0.005344	0.11

Table S4. (continued)

Metabolites	P value	Fold Change	Metabolites	P value	Fold Change
MG(18:1/0:0/0:0)	0.002089	4.40	Phenylalanine	5.91E-16	0.10
Hypoxanthine	1.36E-07	4.36	Acetylcarnosine	3.88E-07	0.10
Piperidine	0.0249	4.30	Betaine	1.00E-12	0.10
Ribulose	0.001168	4.19	Urea	5.63E-10	0.10
Thymidine	4.58E-05	4.12	N-acetyl cadaverine	0.007576	0.09
Uridine	2.97E-19	4.10	ricinoleic acid	2.75E-05	0.09
N-acetylserine	5.28E-09	3.87	Sphingosine	0.002935	0.07
Glutamate	5.28E-09	3.87	adrenic acid	0.000236	0.07
MG(16:1/0:0/0:0)	0.014953	3.53	Proline	4.70E-09	0.06
Glycerol	2.75E-09	3.46	hydroxyarachidonic acid	0.003829	0.06
glutaric acid	5.48E-09	3.38	glycerol-3-phosphate	9.08E-05	0.05
3-hydroxydodecanoic acid	2.07E-22	3.12	Glutamine	2.23E-06	0.04
MG(18:0/0:0/0:0)	0.002584	3.11	phosphoethanolamine	5.35E-07	0.03
Dihydrothymine	8.20E-07	3.05	caprylic acid	6.26E-09	0.03
adipic acid	8.77E-20	2.94	MG(20:4/0:0/0:0)	3.87E-05	0.03
Methyladenine	1.10E-17	2.94	Ornithine	9.28E-16	0.03
Aspartate	7.89E-05	2.76	Taurine	9.25E-19	0.03
4-aminobutyrate	4.90E-11	2.61	Histidine	3.94E-11	0.02
Citrulline	1.15E-08	2.59	Cytosine	0.001714	0.02
Threonine	2.50E-06	2.58	glycerophosphocholine	0.000613	0.01
Putrescine	4.10E-05	2.50			
iminoaspartic acid	0.000117	2.46			

***FC represents fold change of OSCC versus PML, only metabolites with FC values larger than 2.0 or smaller than 0.5 were listed in the table.**

Table S5. Summary of altered metabolic pathways during progression from normal status to premalignant lesion.

Metabolic Pathway	Hits/Total	-LOG$_{10}$(p)	FDR	Impact	Related Metabolites
Aminoacyl-tRNA biosynthesis	6/48	2.95597	0.092966	0	histidine; phenylalanine; arginine; aspartate; lysine; proline
lysine degradation	3/25	2.48745	0.13671	0.14554	cadaverine, piperidine, N-acetyl cadaverine
histidine metabolism	3/16	2.15080	0.19786	0.22131	histidine, N-methyl-histidine, aspartate
glycerophospholipid metabolism	5/36	1.90847	0.21485	0.04684	choline, phosphorylcholine, glycerophosphocholine, phosphoryl ethanolamine, ethanolamine
arginine and proline metabolism	4/38	1.82664	0.21485	0.2678	arginine, proline, putrescine, N-acetyl putrescine
sphingolipid metabolism	3/21	1.81398	0.21485	0.21298	sphinganine, sphingosine, phosphoryl ethanolamine
arginine biosynthesis	2/14	1.31343	0.51491	0.07614	arginine, aspartate

Table S6. Summary of altered metabolic pathways during progression from healthy control to oral squamous cell carcinoma.

Metabolic pathway	**Hits /Total**	**-log$_{10}$(P)**	**FDR**	**Impact**	**Related metabolites**
Aminoacyl-tRNA biosynthesis	13/48	5.72	0.000162	0.16667	histidine, phenylalanine, arginine, glutamate, glutamine, aspartate, serine, lysine, proline, threonine, methionine, leucine/isoleucine, tryptophan
arginine biosynthesis	7/14	5.26	0.000229	0.48223	glutamate, arginine, citrulline, aspartate, ornithine, glutamine, urea
arginine and proline metabolism	11/38	4.36	0.001214	0.51095	arginine, creatine, 4-aminobutanoate, putrescine, spermidine, N-acetyl putrescine, spermine, proline, glutamate, ornithine, phosphocreatine
lysine degradation	7/25	4.24	0.001214	0.31456	lysine, cadaverine, pipecolic acid, piperidine, carnitine, N-acetyl cadaverine, 5-aminopentanoic acid
valine, leucine, and isoleucine metabolism	4/8	3.15	0.011874	0	threonine, leucine/isoleucine, valine
histidine metabolism	5/16	2.74	0.025194	0.34426	glutamate, urocanate, histidine, N-methyl-histidine, aspartate
glycine, serine and threonine metabolism	6/33	1.93	0.14107	0.31291	serine, choline, betaine, phosphoserine, threonine, creatine
glutathione metabolism	5/28	1.64	0.23961	0.03404	glutamate, ornithine, putrescine, spermidine, cadaverine, spermine
beta-alanine metabolism	4/21	1.48	0.27781	0.05597	aspartate, spermine, histidine, spermidine
sphingolipid metabolism	4/21	1.48	0.27781	0.05597	sphingosine, serine, phosphorylethanolamine, phytospinghosine
cysteine and methionine metabolism	4/33	1.36	0.29469	0.22792	serine; methionine, aminobutanoate, phosphoserine
glutamine/glutamate metabolism	2/6	1.34	0.29469	0.5	glutamine, glutamate
purine metabolism	7/65	1.41	0.32372	0.08426	adenosine, AMP, inosine, hypoxanthine, guanosine, urea, allantoin

Table S7. Weight coefficients of the metabolite ions in Lasso regression model.

metabolite	adduct ion	cmz	weight
unknown	---	50.1464	0.0009092
unknown	---	52.2135	0.0111833
unknown	---	54.5165	0.0218527
unknown	---	55.2138	0.0104898
unknown	---	57.487	0.0321527
unknown	---	58.6449	-0.0026
unknown	---	59.6839	0.0211827
unknown	---	65.5846	0.0380753
unknown	---	74.9782	-0.039842
Cadaverine	$[M+H]^+$	103.1227	0.0206961
unknown	---	110.0717	-0.0051
Hypoxanthine	$[M+H]^+$	137.0458	0.0008847
proline	$[M+Na]^+$	138.0525	-0.011193
Taurine	$[M+Na]^+$	148.0035	-0.015395
Proline	$[M+K]^+$	154.0265	-0.027758
creatine	$[M+Na]^+$	154.0587	-0.02682
Glutamine	$[M+Na]^+$	169.0584	-0.012903
Creatine	$[M+K]^+$	170.0323	-0.010836
N6,N6,N6-Trimethyl-L-lysine	$[M+H2O-H]^+$	171.149	0.0384406
unknown	---	178.133	0.0297669
unknown	---	181.5096	0.0164721
unknown	---	182.5849	0.0424328
Cytosine	$[M+2K-H]^+$	187.9574	-0.0558
N-(gamma-Glutamyl)ethanolamine	$[M+H]^+$	191.1026	0.0011292
dimethylarginine	$[M+H]^+$	203.1498	-0.005698
Tyrosine	$[M+Na]^+$	204.0627	0.1502359
unknown	---	212.0396	-0.026289
Vanillylmandelic acid	$[M+Na]^+$	221.0417	-0.043441
Hydroxydodecanoic acid	[M+Na]+	239.1599	0.0036741
Phenylalanine	$[M+2K-H]^+$	241.9998	0.0310015
Dihydrodipicolinate	$[M+2K-H]^+$	245.9537	-0.017957

Table S7 (Continued)

metabolite	adduct ion	cmz	weight
unknown	---	250.1776	-0.00164
N1,N8-Diacetylspermidine	$[M+Na]^+$	252.1674	-0.000022
unknown	---	258.8173	0.032981
Phosphorylcholine	$[M+2K-H]^+$	259.9846	-0.02264
unknown	---	263.9797	-0.03297
unknown	---	279.1586	0.003764
Palmitic acid	$[M+Na]^+$	279.2292	-0.01135
N-Undecanoylglycine	$[M+K]^+$	282.1466	-0.10594
unknown	---	286.3097	-0.0115
unknown	---	286.7811	0.011474
N1,N12-Diacetylspermine	$[M+H]^+$	287.2438	-0.04209
unknown	---	301.1402	0.047266
unknown	---	302.8286	0.039726
unknown	---	303.3072	-0.00754
Oxooctadecanoic acid	$[M+Na]^+$	321.2382	0.023969
Sedoheptulose-phosphate	$[M+K]^+$	329.0034	0.00043
unknown	---	331.3395	-0.01991
unknown	---	337.1034	-0.08058
unknown	---	338.8075	0.017967
Sucrose	$[M+Na]^+$	365.1046	-0.0103
MG(16:1/0:0/0:0)	$[M+K]^+$	367.2235	-0.00222
unknown	---	369.2036	-0.01138
unknown	---	376.2965	-0.01638
Stearoyllactic acid	$[M+Na]^+$	379.2809	-0.03658
Sucrose	$[M+K]^+$	381.0786	-0.02281
Unknown	---	388.7128	0.03562
Unknown	---	407.8493	-0.05132
Unknown	---	409.8474	-0.03127
Unknown	---	421.0684	-0.00395
Unknown	---	423.0652	-0.12462
Unknown	---	514.3815	-0.00025
Intercept	---	---	1.987649

Table S8. The Lasso performance during the 20-fold cross validation.

	Accuracy (%)			
Round	Training set	Testing set	Lambda	DF
1	95.08	90.00	0.025393	49
2	93.99	100.00	0.024885	55
3	93.48	100.00	0.027336	45
4	93.99	90.00	0.027282	50
5	95.08	70.00	0.025166	54
6	94.02	66.67	0.025066	47
7	94.02	88.89	0.027334	50
8	94.02	100.00	0.02733	49
9	94.54	100.00	0.024869	52
10	94.57	100.00	0.020676	56
11	**95.63**	**90.00**	**0.020644**	**62**
12	93.37	75.00	0.030385	37
13	94.02	88.89	0.024979	55
14	91.30	88.89	0.032921	42
15	92.39	88.89	0.030012	45
16	96.72	80.00	0.020864	66
17	94.51	90.91	0.020725	60
18	93.48	77.78	0.02778	48
19	95.11	88.89	0.025127	48
20	93.44	80.00	0.027485	45

Table S9. Prediction performance of the developed Lasso regression model.

	Training dataset **(first batch of 193 cases)**		
Target/Predict	**HC**	**PML**	**OSCC**
HC	62	3	0
PML	0	62	2
OSCC	0	4	60
General accuracy	95.3%		

	Validation dataset **(second batch of 180 cases)**		
Target/Predict	**HC**	**PML**	**OSCC**
HC	52	7	1
PML	0	58	2
OSCC	0	14	46
General accuracy	86.7%		

	Total dataset **(two batch of 373 cases)**		
Target/Predict	**HC**	**PML**	**OSCC**
HC	114	10	1
PML	0	120	4
OSCC	0	18	106
General accuracy	91.2%		

Table S10. Results of ROC analysis for the training and validation datasets

Training	**PML vs HC**	**OSCC vs PML**	**OSCC vs HC**
AUC (CI)	0.9966	0.9968	1.000
Sensitivity	100.0%	93.75%	100.0%
Specificity	98.46%	98.44%	98.46%

Validation	**PML vs HC**	**OSCC vs PML**	**OSCC vs HC**
AUC (CI)	0.9719	0.9169	0.9917
Sensitivity	100.0%	61.67%	90.0%
Specificity	96.67%	98.33%	98.31%

Total	**PML vs HC**	**OSCC vs PML**	**OSCC vs HC**
AUC (CI)	0.9879	0.9627	0.9976
Sensitivity	100.0%	77.42%	95.16%
Specificity	98.13%	99.19%	99.20%

CI: 95% confidence interval

Table S11. The models' performance comparison for the two batches of dataset.

			Training Dataset		Validation Dataset	
Task	**Model**	**Features***	**Accuracy**	**MSE**	**Accuracy**	**MSE**
Classification	ANN	626/626	96.4%	0.1430	90.0%	0.2333
Regression	Lasso	62/626	95.3%	0.1135	86.7%	0.1724
Regression	Quadratic SVM	626/626	92.7%	0.1173	81.7%	0.1667
Classification	Cosine KNN	626/626	92.7%	0.1192	84.4%	0.2722
Regression	Ensemble (Boosted Trees)	626/626	88.1%	0.1346	83.3%	0.1776
Regression	Coarse DT	626/626	85.5%	0.1427	80.0%	0.2804
Classification	Kernal NB	626/626	79.3%	0.2383	55.6%	0.4944

*The internal standard peak was excluded.