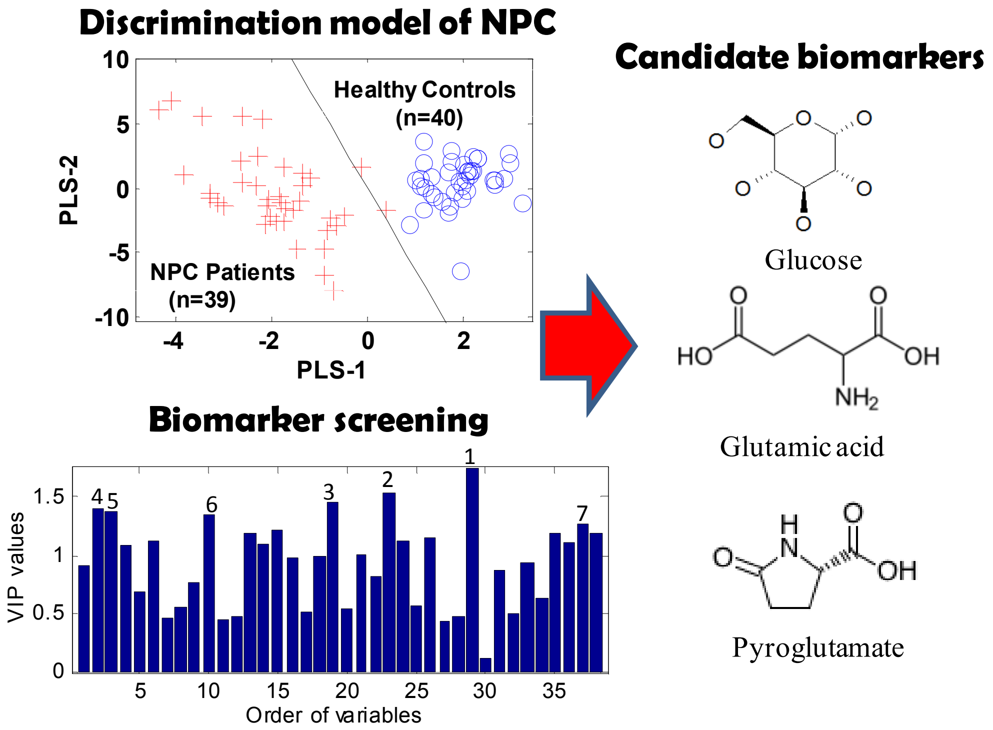
Volume 1 | Number 1 | Jan 2013 | Pages 1–100
RSC Advances
www.rsc.org/advances
ROYAL SOCIETY OF CHEMISTRY

Metabolomic identification of novel biomarkers of nasopharyngeal carcinoma

Lunzhao Yi [a]*, Naiping Dong[c], Shuting Shi[b], Baichuan Deng[d], Yonghuan Yun[b], Zhibiao Yi[e], Yi Zhang[b]*



Metabolomic identification of novel biomarkers of nasopharyngeal carcinoma

Lunzhao Yi [a]*, Naiping Dong[c], Shuting Shi[b], Baichuan Deng[d], Yonghuan Yun[b], Zhibiao Yi[e], Yi Zhang[b*]

[a]Yunnan Food Safety Research Institute, Kunming University of Science and Technology, Kunming, 650500,China

[b]College of Chemistry and Chemical Engineering, Central South University, Changsha, 410083, China

[c]Department of Applied Biology and Chemical Technology, The Hong Kong Polytech nic University, Hong Kong, 999077, China

[d]Department of Chemistry, University of Bergen, Bergen, N-5007, Norway

[e]Dongguan Mathematical and Engineering Academy of Chinese Medicine, GuangZhou University of Chinese Medicine, Dongguan, 523808, China

*Correspondence to: Lunzhao Yi, Yunnan Food safety research institute, Kunming University of Science and Technology, Kunming, 650500, China. Tel.: +86 871 65920302. E-mail address: ylz7910@hotmail.com. Yi Zhang, College of Chemistry and Chemical Engineering, Central South University, Changsha, 410083, China. Tel.: +86 731 88836954, E-mail: yzhangcsu@csu.edu.cn

Abstract

This paper introduces a new identification strategy of novel metabolic biomarkers for nasopharyngeal carcinoma (NPC). We here combined gas chromatography-mass spectrometry (GC-MS) metabolic profiling with three partial least squares-discriminant analysis (PLS-DA) based variable selection methods to screen the metabolic biomarkers of NPC. We found that the variable importance on projection (VIP) method exhibited better efficiency than coefficients β and the loadings plot for the metabolomics data set of 39 NPC patients and 40 healthy controls. In addition, we proved that the area under receiver operating characteristic curve (AUC) was more sensitive than correct rate to evaluate the discrimination ability of classical models. Therefore, three novel candidate biomarkers, glucose, glutamic acid and pyroglutamate were identified with the correct rate was 97.47% and AUC value was 97.40%. Our results suggested that metabolic disorders of NPC mainly reflected in glycolysis and glutamate metabolism; besides, metabolic levels of the related metabolic pathways may affect each other, such as the TCA cycle and lipid metabolism. We believe that the findings of these novel metabolites will be very helpful for early-diagnosis and subsequent pathogenesis research of NPC.

Keywords: Metabolomics; Nasopharyngeal carcinoma; Biomarkers; Variable selection; PLS-DA

1 Introduction

Nasopharyngeal carcinoma (NPC) is a leading cause of cancer death in southern China, where the incidence is 20–40 per 100,000 person-years [1],although it is a rare malignant disease in most parts of the world[2, 3]. NPC is caused by a combination of factors including viral, environmental influences and heredity. Early-diagnosis of NPC is of fundamental importance to prognosis of NPC treatment. Unfortunately, most NPC patients in southern China remain undiagnosed until they present cervical lymph nodes and distant metastasis [4]. A great many researchers are dedicating into new strategies to improve the overall prognosis and reduce morbidity of the NPC patients.

Metabolomics has recently attracted increasing interest in the field of disease diagnosis, pathology, toxicology, and so on, since it is intriguing to be a fast and reproducible method directly reflecting biological events[5-8]. It is well known as a powerful tool to the discovery of biomarkers that may provide additional sensitivity or earlier detection of a disease than classical analytical techniques or histopathology evaluation[5, 9]. A commonly flowchart of metabolomics is the global determination of metabolites followed by disease classification and biomarker screening. Scott et al. counted the papers using classifier approaches published in several journals, such as *Anal. Chem., Anal. Chim. Acta, Metabolomics, et al.* over ten years (2002–2012) [10]. Among all known methods, partial least squares-discriminant analysis (PLS-DA) is the most attractive one in metabolomics research [11-13]. There are several PLS-DA based variable selection methods using to biomarker screening [14], including the loadings plot [15, 16], original coefficients of PLS-DA (β) [17-19] and variable importance on projection (VIP) [20-22] . However, the difficulty for defining the threshold and the problem of different variable combination with the same correct rate cause the complexity of biomarker screening. The selection of efficiency index for class model evaluation is of great importance in biomarker screening.

In this study, we adopted gas chromatography-mass spectrometry (GC-MS) to analyze metabolites of sera samples from 40 healthy donors and 39 newly-diagnosed NPC patients. The flowchart of the study are following: (1) analyze the serum metabolic levels and metabolic characteristics of NPC patients; (2) determine which variable selection method is more suitable for our data set in biomarker screening; (3) determine which index is more efficiency to evaluate the classification ability of a model; and (4) identify a pattern of biomarkers for detection of NPC patients. In addition, the super and sub metabolic pathways of each metabolite were searched and analyzed through KEGG and HMDB data bases, and therefore the alterations of metabolic levels could be correlated with their metabolic pathways. We reported the novel metabolic biomarkers of nasopharyngeal carcinoma, which will be very helpful for NPC diagnosis and further pathogenesis research.

2 Experimental

2.1 Sample collection and Patients

The study was approved by the Human Ethics Committee of Xiangya Hospital, Central South University, and the informed consent was given by each patient for sample collection. In this study, sera samples from 40 healthy volunteers and 39 NPC patients were collected for modeling at the time of diagnosis without any anti-cancer treatment. Age- and gender-matched serum samples from healthy blood donors were used as control group. All serum samples were obtained at February to June 2011 from Xiangya Hospital of Central South University, Hunan, China. The patients' characteristics with respect to age, sex, and ethnic origin were recorded. All investigated patients were uniformly given a routine diagnostic workup comprised of a detailed clinic examination of the head and neck, nasopharyngoscopy, histological and cytological examination of tumor tissue, and radiological imaging examinations (including computed tomography (CT), magnetic resonance imaging (MRI) and ultrasonography). In order to avoid the interferences from post-prandial phase, all sera samples were collected from patients or volunteers fasting at least eight hours. The characteristics of NPC patients and controls were shown in table 1.

Insert Table 1

2.2 Chemicals and reagents

BSTFA+1%TMCS (N,O-Bis(trimethylsilyl) trifluoroacetamide with 1% trimethylchorosilane, for GC) (>99.0% purity), pyridine(>99.8% purity) and methoxyamine hydrochloride (>98% purity), and the other 25 chemical standards of metabolites (shown in table 2) were purchased from Sigma-Aldrich (St. Louis, MO, USA). Methanol is analytical grade and purchased from the Hanbang Chemical Corporation (Zhenjiang, China).

2.3 GC-MS data acquisition

Blood sample (4 ml) was allowed to clot at 4 °C and was centrifuged at 2000 g for 20 min. Sera were collected, aliquoted, and stored at -80 °C until the analysis was carried out. Briefly, each 100 µl serum sample was mixed with 350 µl methanol, and 50 µl heptadecanoic acid (dissolved in methanol at a concentration of 1 mg/ml) was added as an internal standard. After vigorously vortexing for 1 min, the mixture was centrifuged at 16000 rpm for 10 min at 4 °C. The supernatant (400 µl) was transferred to a 5 ml glass centrifugation tube and evaporated to dryness under N_2 gas. Then, 70 µl of methoxyamine hydrochloride solution (20 mg/ml in pyridine) was added into the residue and incubated for 60 min at 70 °C. After methoximation, 100µl of BSTFA derivitization agent was added into the residue and incubated for another 50 min at 70 °C. The final solution was used for GC-MS analysis.

All GC-MS analyses were performed by a gas chromatography instrument (Shimadzu GC2010A, Kyoto, Japan) coupled to a mass spectrometer (GC-MS-QP2010) with a constant flow rate of helium carrier gas at 1.0 ml/min. For each sample, 1.0 µl was injected into a DB-5ms capillary column (30 m×0.25 mm i.d., film thickness is 0.25

μm) at a split ratio of 1:10. The column temperature was initially maintained at 70 °C for 4 min, and then increased at a rate of 8 °C/min from 70 to 300°C and held for 3 min. The total GC run time was 35.75 min. Mass conditions were maintained as followed: ionization voltage, 70 eV; ion source temperature, 200 °C; interface temperature, 250 °C; full scan mode in the 35–800 amu mass ranges with 0.2 s scan velocity; detector voltage, 0.9 kV.

2.4 GC-MS data processing

All GC-MS data, including retention characteristics, peak intensities, and integrated mass spectra, of each serum sample were used for the analysis. Firstly, the automated mass-spectral deconvolution and identification system (AMDIS software, National Institute of Standards and Technology, Gaithersburg, MD) was employed to support peak finding and deconvolution. Using NIST Mass Spectral Search Program Version 2.0 and the characteristic ions, tentative identification of structures of peaks-of-interest was supported by similarity search of the NIST/EPA/NIH Mass Spectra Library (NIST05), which contained 190,825 EI spectra for 163,198 compounds. 38 metabolites were considered to be the main endogenous metabolites.25 metabolites were identified by their corresponding chemical standards. The peak areas of metabolites were compared with that of the internal standards to provide the semi-quantitative level for the metabolites. The peak areas were extracted using our custom scripts to generate a data matrix, in which the rows represent the samples and the columns correspond to peak/area ratios to the internal standard in the same chromatogram. The size of the matrix is 79×38.

2.5 Statistical analysis

All datasets were autoscaled before PLS-DA. Data matrix of relative peak areas generated from metabolic profiles were analyzed by PLS-DA, in order to establish any “groupings” with respect to NPC patients and healthy controls. 10-fold cross validation was employed to select the optimal number of latent variables and evaluate the predictive ability of PLS-DA model. Permutation test were employed to evaluate the reliability of the class model and calculated 5000 times. In addition, two indexes, correct rate and the area under receiver operating characteristic curve (AUC), were compared to evaluate the classification ability of a model.

After the discrimination model was established by PLS-DA, the variable selection is carried out to identify the novel biomarkers. The loadings plot, original coefficients of PLS-DA (β) and variable importance on projection (VIP) were employed and compared. The three methods are commonly used in metabolomics.

The loadings plot: generally, the loadings plot indicates the influence of original variables on the corresponding scores. So, if the scores plot can discriminate the different classes of samples, the loadings plot can partly express the influence of variables on separation between classes. These variables having the greatest influence on the scores plot are furthest away from the main cluster of variables.

Original coefficients of PLS-DA (β): the vector of β is the coefficients of the PLS transformed equation between the discriminant equation expressed by latent variables obtained by PLS and that expressed by the original variables. It is a single measure of association between each variable and the response. For the autoscaled data, the absolute value of β can render the influence of the corresponding variables on the separation between sample classes. The higher the absolute value of β is, the more the influence of corresponding variable is.

Variable importance on projection (VIP): the idea behind this measure is to accumulate the importance of each variable *j* being reflected by ***w*** from each latent variables (scores). ***w*** is the weight of PLS analysis. The VIP measure v_j is defined as

$$v_j = \sqrt{p\sum_{a=1}^{A}\left[SS_a(w_{aj}/\|w_a\|^2)\right]/\sum_{a=1}^{A}SS_a}$$

where p is the a*th* loading, SS_a is the sum of squares explained by the *a*th latent variable (score). Hence, the v_j weights is a measure of the contribution of each variable according to the variance explained by each PLS latent variable where $(w_{aj}/ \| w_a \|)^2$ represents the importance of the *j*th variable [14]. The higher the value of VIP is, the more the influence of corresponding variable is.

All programs of PLS-DA and other methods were coded in MATLAB 2010 for Windows and all calculations were performed on an Intel Core i7 processor based personal computer with 16G RAM memories.

3 Results and discussion

3.1Metabolic profiling

38 metabolites, involved in the metabolic processes of amino acid, carbohydrate, energy, lipid, organic acid and urea, were qualitatively and quantitatively analyzed in details, shown in table 2. For each metabolite, the statistical significance of the differences between NPC patients and controls was calculated separately by Mann–Whitney U test. Serum levels of 12 metabolites increased strikingly in NPC patients compared with controls, while 7 metabolites significantly decreased (Mann–Whitney U test p<0.05 with a signed t value of ‘‘1’’). For NPC patients, mean level of lactate, an end product of glycolsis, increased by 42%. Mean level of malic acid, an intermediate in the tricarboxylic acid cycle (TCA cycle), also increased by 50%.. Mean level of glutamic acid, a key compound in cellular metabolism, increased by 221%. Palmitic acid (C16:0), stearic acid (C18:0) and cholesterol increased by 25%, 39% and 23%, respectively. They all belong to lipid group. Mean levels of three

unsaturated fatty acids, linoleic acid (C18:2n6), oleic acid (C18:1n9) and arachidonic acid (C20:4n6) decreased by 19%, 19% and 32% for NPC compared with controls, respectively. Galactose and glucose levels decreased 24% and 40%, respectively. These findings suggested that serum metabolic disorders appeared mainly in glutamate, glycolysis, krebs cycle and lipid metabolism for NPC patients.

Insert Table 2

3.2 Discrimination model between NPC patients and controls

PLS-DA was employed to establish a discrimination model between NPC patients and healthy controls. The autoscaled data set of 38 metabolites was used as input data. 10-fold cross validation was applied to select the optimal number of latent variables. A 2-dimensional PLS-DA model constructed by the first two latent variables (PLS-1 and PLS-2) was obtained (Fig.1 (A)). In addition, the reliability and predictive ability of the model was evaluated by permutation test (Fig.1 (B)) and 10-fold cross validation. The data set was permutated for 5000 times. The frequency of correct rates for the 5000 permutated models is a normal distribution with mean value near 50% (Fig.1 (B)), which guarantee the reliability of the established discrimination model. The NPC and control samples were separated clearly by the discriminant line (Fig. 1(A)) with a total correct rate of 97.47%. The AUC is 97.44%. The correct rates of 10-fold cross validation for controls, NPCs and the total were 100% (40/40), 94.87% (37/39) and 97.47% (77/79), respectively. The AUC is 96.86%. These results indicated that the established PLS-DA model is reliable and with good classification ability to discriminate NPC patients from healthy controls.

Insert Figure 1

3.3 Identification of Candidate Biomarkers for NPC

After the metabolic discrimination model was established by PLS-DA, variable selection was carried out to identify the candidate biomarkers of NPC. Three variable selection methods were employed and compared, including the loadings plot, original coefficients of PLS-DA (β) and VIP. Though candidate biomarkers selected by these three variable selection methods are not the same, shown in Fig.1 (C), (D) and (E), there are some common metabolites. Two metabolites, glutamic acid (23) and glucose (29), were identified as the first and second important metabolites by all the three methods. A PLS-DA model established by the two metabolites had good classification ability. Correct rate of 10-fold cross validation is 91.14% (Table 3). The AUC value is 97.24% (Table 3). The results indicated that glutamic acid and glucose are very important metabolites for NPC metabolic disorders, representing many metabolic characteristics of this disease.

Insert Table 3

In addition, combination effect of variables was taken into account in this study. Classification ability of different variable combinations was compared in order to select the best biomarker pattern and help us to define the threshold of variable selection. The number of variables varied from one to seven. For VIP method, the best result of correct rate and AUC of 10-fold cross validation (correct rate: 97.47%, AUC: 97.40%) was obtained when the number of variables is three, shown in Fig.1 (F) and (G), Table 3. The selected metabolites are pyroglutamate (19), glutamic acid (23) and glucose (29). For coefficients β, correct rate of the model established by the first three metabolites is 92.41%, AUC value is 96.79%. Until the number of variables is seven, correct rate is as good as the three metabolites selected by VIP (correct rate: 97.47%). In fact, there are four different variable combinations with the same correct rate (correct rate: 97.47%). It is very difficult to decide which variable combination is the best based on the results of correct rates. For AUC value, only one variable combination has the best result (glucose, glutamic acid and pyroglutamate, AUC: 97.40%), which is select by VIP. It seems that the value of AUC is more sensitive to evaluate the discrimination ability of a model for our data set. In this study, the combination of metabolites identified by VIP method gets the best discrimination results evaluated by both AUC value and correct rate. We suggested that VIP method is more effective than coefficients β and the loadings plot for our data set.

In the loadings plot, the projection points of variables are scattered for the autoscaled data set (Fig.1 (E)). Though the three metabolites, pyroglutamate (19), glutamic acid (23) and glucose (29), could be screened by this method, it is subjective and easy to be disturbed by other metabolites.

3.4 Associations between identified biomarkers and NPC

In this study, three candidate biomarkers, glucose, glutamic acid and pyroglutamate were identified, mainly belonging to two metabolic pathways, glycolysis and glutamate metabolism.

Glucose is identified as the most important metabolite for NPC by the three variable selection methods. For NPC patients, mean level of glucose decreased by 40% compared with controls (Table 2), decreased by 51% in our former research [12]. The correct rate of the classification model established only by glucose was 88.61% (AUC: 91.25%), which indicated the good classification ability of glucose (Table 3). Glucose is a primary source of energy for living organisms. It is reported that in tumor cells, glucose utilisation is greatly enhanced compared with that of normal tissue [23]. Unlike their normal counterparts, tumor cells preferentially use enhanced aerobic glycolysis for energy metabolism, a phenomenon first described by Otto Warburg in 1925 and known as the Warburg effect [24]. This shift toward increased glycolytic flux allows tumor cells to produce sufficient ATP to fulfill metabolic demands and leads to increased glucose consumption, decreased oxidative phosphorylation, and increased lactate production [25]. In this study, the alterations of glucose (decreased by 40%) and lactate (increased by 42%) levels in serum are consistent with the results of reported researches on tumor tissues and cells. In addition, there is another metabolite 1,5-anhydro-sorbitol (1,5-AG) related with the alterations of glucose level. 1,5-AG is a metabolite used to identify glycemic variability in people with diabetes. It is reported that 1,5-AG decreases during times of hyperglycemia above 180 mg/dL, and returns to normal levels after approximately 2 weeks in the absence of hyperglycemia [26]. In this study, serum 1,5-AG level increased by 43%, while glucose level decreased. It suggested that a biological process opposite to hyperglycemia may happen for NPC. However, the reason of these alterations is not clear and needs our further research.

Glutamic acid is the second important metabolite selected by VIP. Recently, a paper published in *Nature* reported that glutamine (Gln) supports pancreatic cancer growth through a KRAS-regulated metabolic pathway. Consistent with this observation, glutamate (glutamic acid, Glu) is able to support growth in Gln-free conditions [27]. In our study, serum level of glutamic acid (Glu), a degradation product of Gln, increased obviously for NPC patients, by 221% compared with controls (Table 2). It seems that disorders of glutamate metabolism are serious for NPC. In addition, Glu could be converted into a-ketoglutarate to replenish the TCA cycle through two mechanisms [28]. Serum levels of malic acid, a metabolite in TCA cycle, increased by 50% for NPC patients. The results suggested that some metabolic pathways may exist to link glutamate metabolism and TCA cycle for NPC metabolic disorders.

Pyroglutamic acid is a cyclized derivative of Glu. Abnormal blood level may be associated with problems of glutamine or glutathione metabolism. Serum level of pyroglutamate for NPC decreased by 24% compared with controls (Table 2), decreased by 43% for another groups of NPC sera samples in our former research [12]. In the former study, pyroglutamate was not identified as one of the marker metabolites contributing to the discrimination between NPC and controls, because of differences of samples and the limitation of data processing method. However, it is found that levels of pyroglutamate increased obviously three months after treated with the standard radiotherapy [12]. In this study, pyroglutamate is identified as one of the candidate biomarkers for NPC with the help of VIP.

4 Conclusion

In summary, this study demonstrated a convincing strategy for novel metabolic biomarkers identification by combining GC-MS metabolic profiling with variable selection methods based on PLS-DA. This protocol has been successfully applied to metabolomics research of nasopharyngeal carcinoma and three candidate biomarkers, glucose, glutamic acid and pyroglutamate were identified in this study. It needs to be emphasized that the efficiency of VIP method is much higher than coefficients β and the loadings plot for our data set. In addition, two indexes, correct rate and AUC value of ROC curve, were employed to evaluate the discrimination ability of a class model, while the value of AUC exhibit better sensitivity. Our results suggest that metabolic disorders of nasopharyngeal carcinoma are mainly reflected in glycolysis and glutamate metabolism. We also suggest that the metabolic levels of the related metabolic pathways may affect each other, such as the TCA cycle and lipid metabolism. We here believe that the findings of these novel metabolites will be very helpful for diagnosis and further pathogenesis research of NPC.

Acknowledgements

This work was supported financially by National Nature Foundation Committee of

P.R. China (No.21465016, No.21105129, No.21473257), Science and Technological Program for Dongguan's Higher Education, Science and Research, and Health Care Institutions (2012108102032).

References

1. E. T. Chang and H. O. Adami, *Cancer Epidemiology Biomarkers & Prevention*, 2006, 15, 1765-1777.
2. A. T. C. Chan, V. Gregoire, J. L. Lefebvre, L. Licitra, E. Felip and E.-E.-E. G. Working, *Annals of Oncology*, 2010, 21, v187-v189.
3. C. de Martel, J. Ferlay, S. Franceschi, J. Vignat, F. Bray, D. Forman and M. Plummer, *Lancet Oncology*, 2012, 13.
4. A. M. Mackie, J. B. Epstein, J. S. Y. Wu and P. Stevenson-Moore, *Oral Oncology*, 2000, 36, 397-403.
5. M. Tomita and K. Kami, *Science*, 2012, 336, 990-991.
6. J. K. Nicholson and J. C. Lindon, *Nature*, 2008, 455, 1054-1056.
7. S. Moco, R. J. Bino, R. C. H. De Vos and J. Vervoort, *Trac-Trends in Analytical Chemistry*, 2007, 26, 855-866.
8. J. Gillard, J. Frenkel, V. Devos, K. Sabbe, C. Paul, M. Rempt, D. Inze, G. Pohnert, M. Vuylsteke and W. Vyverman, *Angewandte Chemie-International Edition*, 2013, 52, 854-857.
9. A. Sreekumar, L. M. Poisson, T. M. Rajendiran, A. P. Khan, Q. Cao, J. Yu, B. Laxman, R. Mehra, R. J. Lonigro, Y. Li, M. K. Nyati, A. Ahsan, S. Kalyana-Sundaram, B. Han, X. Cao, J. Byun, G. S. Omenn, D. Ghosh, S. Pennathur, D. C. Alexander, A. Berger, J. R. Shuster, J. T. Wei, S. Varambally, C. Beecher and A. M. Chinnaiyan, *Nature*, 2009, 457, 910-914.
10. I. M. Scott, W. Lin, M. Liakata, J. E. Wood, C. P. Vermeer, D. Allaway, J. L. Ward, J. Draper, M. H. Beale, D. I. Corol, J. M. Baker and R. D. King, *Analytica Chimica Acta*, 2013, 801, 22-33.
11. B. J. Blaise, A. Gouel-Cheron, B. Floccard, G. Monneret and B. Allaouchiche, *Analytical Chemistry*, 2013, 85, 10850-10855.
12. L. Yi, C. Song, Z. Hu, L. Yang, L. Xiao, B. Yi, W. Jiang, Y. Cao and L. Sun, *Metabolomics*, 2013, DOI: 10.1007/s11306-013-0606-x, 1-12.
13. J. A. Westerhuis, H. C. J. Hoefsloot, S. Smit, D. J. Vis, A. K. Smilde, E. J. J. van Velzen, J. P. M. van Duijnhoven and F. A. van Dorsten, *Metabolomics*, 2008, 4, 81-89.
14. T. Mehmood, K. H. Liland, L. Snipen and S. Saebo, *Chemometrics and Intelligent Laboratory Systems*, 2012, 118, 62-69.
15. Z. Huang, Y. Chen, W. Hang, Y. Gao, L. Lin, D. Y. Li, J. Xing and X. Yan, *Metabolomics*, 2013, 9, 119-129.
16. D. Paris, D. Melck, M. Stocchero, O. D'Apolito, R. Calemma, G. Castello, F. Izzo, G. Palmieri, G. Corso and A. Motta, *Metabolomics*, 2010, 6, 405-416.
17. L. Z. Yi, J. He, Y. Z. Liang, D. L. Yuan and F. T. Chau, *Febs Letters*, 2006, 580, 6837-6845.
18. J. T. Brindle, H. Antti, E. Holmes, G. Tranter, J. K. Nicholson, H. W. L. Bethell, S. Clarke, P. M. Schofield, E. McKilligin, D. E. Mosedale and D. J. Grainger, *Nature Medicine*, 2002, 8, 1439-1444.
19. P. Zheng, Y. D. Wei, G. E. Yao, G. P. Ren, J. Guo, C. J. Zhou, J. J. Zhong, D. Cao, L. K. Zhou and P. Xie, *Metabolomics*, 2013, 9, 800-808.
20. J. Yang, X. J. Zhao, X. L. Liu, C. Wang, P. Gao, J. S. Wang, L. J. Li, J. R. Gu, S. L. Yang and G. W. Xu, *Journal of Proteome Research*, 2006, 5, 554-561.
21. P. Yin, X. Zhao, Q. Li, J. Wang, J. Li and G. Xu, *Journal of Proteome Research*, 2006, 5, 2135-2143.

22. J. Xu, Y. H. Chen, R. P. Zhang, Y. M. Song, J. Z. Cao, N. Bi, J. B. Wang, J. M. He, J. F. Bai, L. J. Dong, L. H. Wang, Q. M. Zhan and Z. Abliz, *Molecular & Cellular Proteomics*, 2013, 12, 1306-1318.

23. T.-C. Yen, Y.-C. Chang, S.-C. Chan, J.-C. Chang, C.-H. Hsu, K.-J. Lin, W.-J. Lin, Y.-K. Fu and S.-H. Ng, *Eur J Nucl Med Mol Imaging*, 2005, 32, 541-548.

24. O. Warburg, *Science*, 1956, 123, 309-314.

25. E. Noch and K. Khalili, *Molecular cancer therapeutics*, 2012, 11, 14-23.

26. T. Yamanouchi, N. Ogata, T. Tagaya, T. Kawasaki, N. Sekino, H. Funato, I. Akaoka and H. Miyashita, *The Lancet*, 1996, 347, 1514-1518.

27. J. Son, C. A. Lyssiotis, H. Ying, X. Wang, S. Hua, M. Ligorio, R. M. Perera, C. R. Ferrone, E. Mullarky and N. Shyh-Chang, *Nature*, 2013.

28. M. G. Vander Heiden, L. C. Cantley and C. B. Thompson, *Science*, 2009, 324, 1029-1033.

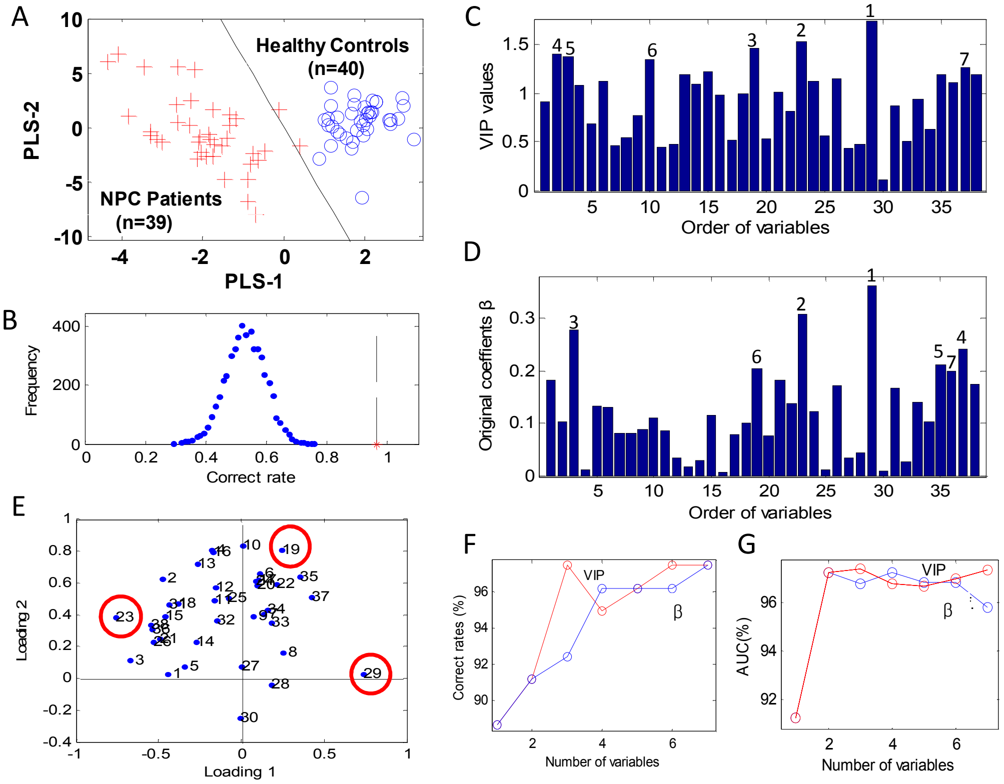


Figure 1: Identification of candidate biomarkers for NPC. (A) PLS-DA model for discrimination between NPC patients and healthy controls. (B) Distribution of 10-fold cross validation correct rates. The asterisk point is the error for current model, and the blue points are the distribution of 5000 times permuted 10-fold cross validation correct rates. (C) VIP value of each metabolite. (D) Original coefficients β of 38 metabolites. (E) The loadings plot. The correct rates (F) and the AUC values (G) of the PLS-DA models of different combinations of variables. The selection of variables was performed according to their value of VIP or β. The first one was the variable with the highest VIP or β value. The second combination was the first one plus the second one, then, the first three, and so on. The correct rate and AUC value was obtained from the 10-fold cross validation. The red and blue lines indicate variables selected by VIP and β, respectively.

Table 1 Characteristics of NPC patients and controls

Characteristics	NPC patients	Healthy controls
No. of subjects	39	40
Race	Han	Han
Age (median)	49 years	41 years
Gender(％men)	56％(22/39)	45％(18/40)
UICC[a] stage(2003)		
Ⅰ	8% (3/39)	-
ⅡA/ⅡB	41% (16/39)	-
Ⅲ	41% (16/39)	-
ⅣA	8% (3/39)	-
ⅣB	2% (1/39)	-
ⅣC	0%	-
KPS score		
≥ 80	80% (31/39)	-
60-80	8% (3/39)	-
30-60	2% (1/39)	-
≤ 30	10% (4/39)	-

Table 2 Qualitative and quantitative analysis of metabolic profiles of healthy controls and NPC patients

No.	Super pathway	Sub pathway	Biochemical name	Relative quantity		t	p	KEGG	HMDB
				Controls (n=40)	NPCs (n=39)				
4	Amino acid	Alanine and aspartate metabolism	alanine*	0.124±0.034	0.127±0.047	0	0.69	C00041	HMDB00161
5		Glycine,serine and threonine metabolism	sarcosine	0.118±0.047	0.147±0.050	1↑	0.008	C00213	HMDB00271
6			glycine*	0.063±0.029	0.050±0.032	0	0.06	C00037	HMDB00123
15			glycerate	0.011±0.004	0.017±0.014	1↑	0.005	C00258	HMDB00139
16			serine*	0.059±0.020	0.061±0.034	0	0.66	C00065	HMDB00187
17			threonine*	0.056±0.021	0.050±0.022	0	0.21	C00188	HMDB00167
10		Valine, leucine and isoleucine metabolism	valine*	0.092±0.025	0.084±0.032	0	0.19	C00183	HMDB00883
12			isoleucine*	0.025±0.010	0.027±0.011	0	0.39	C00407	HMDB00172
13		Urea cycle; arginine-, proline-, metabolism	proline*	0.050±0.017	0.055±0.028	0	0.31	C00148	HMDB00162
20			trans-4-hydroxyproline	0.007±0.004	0.006±0.005	0	0.24	C01157	HMDB00725
19		Glutamate metabolism	pyroglutamate *	0.160±0.042	0.122±0.060	1↓	0.001	C01879	HMDB00267
23			glutamic acid*	0.014±0.007	0.045±0.023	1↑	1.16E-11	C00064	HMDB00148
22		Creatine metabolism	creatinine enol	0.013±0.005	0.010±0.006	1↓	0.02	C00791	HMDB00562
24		Phenylalanine & tyrosine metabolism	phenylalanine*	0.023±0.016	0.018±0.009	0	0.09	C00079	HMDB00159
34		Tryptophan metabolism	tryptophan	0.017±0.005	0.015±0.007	0	0.09	C00078	HMDB00929
2	Carbohydrate	Glycolysis, gluconeogenesis,	lactate*	1.083±0.327	1.533±0.978	1↑	0.007	C00186	HMDB00190

29		pyruvate metabolism	glucose*	4.152±0.433	2.480±1.024	1↓	1.32E-14	C00031	HMDB00122
26		Hexoses	1,5-anhydro-sorbitol*	0.097±0.038	0.139±0.055	1↑	1.67E-04	-	HMDB02712
27		Fructose, mannose, galactose, starch, and sucrose metabolism	fructose*	0.027±0.012	0.029±0.015	0	0.67	C00095	HMDB00660
28			galactose*	0.029±0.006	0.022±0.011	1↓	0.002	C01582	HMDB00143
30			mannose	0.038±0.026	0.040±0.076	0	0.89	C00159	HMDB00169
14	Energy	Krebs cycle	succinate	0.004±0.002	0.004±0.001	0	0.24	C00042	HMDB00254
18			malic acid*	0.002±0.002	0.003±0.002	1↑	0.02	C00149	HMDB00156
25			citric acid*	0.020±0.009	0.021±0.017	0	0.911	C00158	HMDB00094
31	Lipid	Long chain fatty acid	palmitic acid (C16:0)*	0.163±0.041	0.204±0.060	1↑	7.42E-04	C00249	HMDB00220
33		Long chain fatty acid Inositol metabolism	oleic acid (C18:1n9)*	0.192±0.066	0.156±0.073	1↓	0.03	C00712	HMDB00207
36			stearic acid (C18:0)*	0.070±0.023	0.097±0.030	1↑	2.68E-05	C01530	HMDB00827
37			arachidonic acid (C22:4n6)*	0.031±0.010	0.021±0.009	1↓	1.57E-05	C00219	HMDB01043
32			myo-inositol	0.018±0.008	0.019±0.005	0	0.45	C00137	HMDB00211
35		Essential fatty acid	linoleic acid(C18:2n6)*	0.133±0.029	0.108±0.030	1↓	2.29E-04	C01595	HMDB00673
38		Sterol/Steroid	cholesterol*	0.349±0.050	0.428±0.111	1↑	1.09E-04	C00187	HMDB00067
1	Organic acid	Dicarboxylate	Oxalic acid	0.027±0.010	0.036±0.010	1↑	2.78E-04	C00209	HMDB02329
3		Short-chain Hydroxy Acids	Tartronic acid	0.007±0.003	0.012±0.004	1↑	4.39E-09	-	HMDB35227
7		Short-chain Hydroxy Acids Ascorbate and aldarate metabolism	á-Hydroxy butyrate	0.016±0.006	0.014±0.008	0	0.18	C05984	HMDB00008
8			â-Hydroxy butyrate*	0.031±0.033	0.019±0.028	0	0.09	C01089	HMDB00357
9			á-Hydroxyisovaleric acid	0.005±0.003	0.005±0.002	0	0.21	-	HMDB00407

21			2,3,4-trihydroxybutyrate	0.004±0.003	0.006±0.002	1↑	1.46E-04	C01620	HMDB00943
11	Ureas	Arginine and proline metabolism	Urea	0.671±0.300	0.769±0.262	0	0.13	C00086	HMDB00294

38 data are presented as mean ± SD. t is the Mann–Whitney U test results between NPC patients and controls; A p value of <0.05 is considered statistically significant and signed t value is ‘‘1,’’ otherwise ‘‘0.’’ The number of metabolite is listed according to their retention time.* Identified by standard substances.

Table 3 Recognition and predictive ability [a] of the PLS models established by selected metabolites

NoM			Recognition ability	Predictive ability
1(A and B)	glucose (29)	Sensitivity	82.05%	79.49%
		Specificity	97.50%	97.50%
		Correct rate	89.87%	88.61%
		AUC	91.44%	91.25%
2 (A and B)	glutamic acid (23) glucose (29)	Sensitivity	82.05%	82.05%
		Specificity	100%	100%
		Correct rate	91.14%	91.14%
		AUC	97.37%	97.24%
3 (A)	pyroglutamate (19) glutamic acid (23) glucose (29)	Sensitivity	97.44%	94.87%
		Specificity	100%	100%
		Correct rate	98.73%	97.47%
		AUC	97.44%	97.40%
3(B)	Tartronic acid (3) glutamic acid (23) glucose (29)	Sensitivity	89.74%	87.18%
		Specificity	100%	97.50%
		Correct rate	94.94%	92.41%
		AUC	96.96%	96.79%
4 (A)	lactate (2), pyroglutamate (19) glutamic acid (23) glucose (29)	Sensitivity	94.87%	95.00%
		Specificity	97.50%	94.87%
		Correct rate	96.20%	94.94%
		AUC	96.92%	96.79%
4 (B)	Tartronic acid (3) glutamic acid (23) glucose (29) Arachidonic acid (37)	Sensitivity	92.31%	92.31%
		Specificity	100%	100%
		Correct rate	96.20%	96.20%
		AUC	97.44%	97.20%
5 (A)	lactate (2), tartronic acid (3), pyroglutamate (19), glutamic acid (23), glucose (29)	Sensitivity	92.31%	92.31%
		Specificity	100%	100%
		Correct rate	96.20%	96.20%
		AUC	96.83%	96.67%
5 (B)	Tartronic acid (3) glutamic acid (23) glucose (29), Linoleic acid (35) Arachidonic acid (37)	Sensitivity	94.87%	94.87%
		Specificity	97.50%	97.50%
		Correct rate	96.20%	96.20%
		AUC	96.92%	96.83%
6 (A)	lactate (2), tartronic acid (3), norvaline (10), pyroglutamate (19), glutamic acid (23), glucose (29)	Sensitivity	92.31%	94.87%
		Specificity	100%	100%
		Correct rate	96.20%	97.47%
		AUC	97.21%	96.99%
6 (B)	Tartronic acid (3), Pyroglutamate (19) glutamic acid (23) glucose (29), Linoleic acid (35) Arachidonic acid (37)	Sensitivity	97.44%	94.87%
		Specificity	97.50%	97.50%
		Correct rate	97.47%	96.20%
		AUC	97.15%	96.83%

7 (A)	lactate (2), tartronic acid (3), norvaline (10), pyroglutamate (19), glutamic acid (23), glucose (29), arachidonic acid (37)	Sensitivity	94.87%	94.87%
		Specificity	100%	100%
		Correct rate	97.47%	97.47%
		AUC	97.37%	97.31%
7 (B)	Tartronic acid (3), Pyroglutamate (19) glutamic acid (23) glucose (29), Linoleic acid (35), Stearic acid (36), Arachidonic acid (37)	Sensitivity	97.44%	97.44%
		Specificity	97.50%	97.50%
		Correct rate	97.47%	97.47%
		AUC	96.47%	95.77%
5 (C)	tartronic acid (3), pyroglutamate (19), glutamic acid (23), glucose (29), arachidonic acid (37)	Sensitivity	94.87%	92.31%
		Specificity	100%	100%
		Correct rate	97.47%	96.20%
		AUC	97.37%	97.28%

[a] NoM: number of metabolites. Recognition ability is the correct classification of the training. Prediction ability is the rate of the correct classification of the 10-fold cross validation. Sensitivity is the number of true positives classified as positive (patients). Specificity is the number of true negative classified as negative (healthy controls). A: metabolites selected by VIP; B: metabolites selected by original coefficients (β); C: common metabolites selected by VIP and β.