Supplementary information

Supporting text

1. Multivariate statistical analysis for samples in the Hangzhou, Beijing and City of Hope cohorts.

In the supporting figure 3, PCA scores plots for samples in the Hangzhou, Beijing and City of Hope cohorts were demonstrated. From these three PCA scores plots, separation trend between CRC tissues and adjacent non-tumor tissues can be observed. The parameters for these models were: 8 components with $R^2_{X,cum} = 0.556$ for samples in Hangzhou cohort, 5 components with $R^2_{X,cum} = 0.619$ for samples in Beijing cohort and 6 components with $R^2_{X,cum} = 0.66$ for samples in City of Hope cohort.

In the supporting figure 4, OPLS-DA scores plots for samples in the Hangzhou, Beijing and City of Hope cohorts were demonstrated. From these three OPLS-DA scores plots, good separations between CRC tissues and adjacent non-tumor tissues were obtained. The parameters for these models were: one predictive component and two orthogonal components ($R^2_{X,cum}=0.364$, $R^2_{Y,cum}=0.704$, $Q^2_{cum}=0.542$) for samples in Hangzhou cohort, one predictive component and two orthogonal components ($R^2_{X,cum}=0.428$, $R^2_{Y,cum}=0.802$, $Q^2_{cum}=0.532$) for samples in Beijing cohort and one predictive component ($R^2_{X,cum}=0.212$, $R^2_{Y,cum}=0.294$, $Q^2_{cum}=0.152$) for samples in City of Hope cohort. The VIP values for those 15 core metabolites were listed in the Supporting table 1.

2. Diagnostic value of “core” metabolites for gastric cardiac cancer
Using the same analytical platform and protocol, we previously analyzed tumor tissue samples collected from gastric cardiac cancer patients (GCC, 40 paired tumor and adjacent non-tumor tissue samples) to investigate the dysregulated glucose metabolism, the data of which was published in Mol. Cell Proteomics (2010, 9(12): 2617-28) (1). These patients included 37 male and 3 female GCC patients with median age of 59 (42-76). Their pathological stages (TNM stage) were characterized as: 3 at stage I, 17 at stage II, and 20 at stage III.

To make the data comparable with the CRC data, the data was normalized to the mean value of corresponding adjacent non-tumor samples. A binary logical regression with the 15 metabolites was performed with CRC samples (and adjacent non-tumor) collected from Shanghai, which generate a function for Odds ratios. Based on 15 “core” metabolites, a function for Odds calculation was obtained using CRC samples from Shanghai with binary logistic regression. The Function for Odds is:

\[
\text{In(Odds)} = 1.839 \times 2\text{-aminobutyrate} + (1.700) \times 5\text{-oxoproline} + 0.568 \times \text{Aspartate} + 1.201 \times \text{Beta-alanine} + 0.996 \times \text{Cysteine} + (-2.724) \times \text{Glutamate} + (-2.469) \times \text{Glycerol} + 3.501 \times \text{Hypoxanthine} + (-0.308) \times \text{Lactate} + (-4.664) \times \text{Myo-inositol} + (-2.363) \times \text{Myristate} + 1.253 \times \text{Palmitoleate} + 0.166 \times \text{Putrescine} + (-0.045) \times \text{Uracil} + 1.502 \times \text{Kynurenine} + (-2.575).
\]

The probability was calculated with the function of probability = Odds/(1+Ossd).

With the same function listed above, the Odds ratios and the probabilities were calculated for all the samples including 4 batches of CRC samples, and gastric cardiac cancer samples. With a cut-off of 0.5, most of CRC samples of all the four batches were scattered above the cut-off line, while most of the adjacent non-tumor samples collected from CRC patients were scattered below the cut-off line (Supporting figure 3B). For gastric cardiac cancer, no clear separation tendency can be observed (Supporting figure 3B). The sensitivity and specificity for the sample were: 0.953 and 0.873 for CRC
samples collected from Shanghai, 0.692 and 0.846 for CRC samples collected from Hangzhou, 0.956 and 0.782 for CRC samples collected from Beijing, 0.650 and 0.842 for CRC samples collected from City of Hope, 0.350 and 0.775 for GCC samples, respectively. The low sensitivity for identifying GCC tumor tissue samples from non-tumor tissues may indicate the specificity of the “core” metabolites pattern on discriminating CRC.

3. Prognostic values of selected metabolite markers for CRC

To simplify metabolic patterns used to predict the prognostic value for those CRC patients, we tried the ratios of two metabolites to predict the outcomes after surgical treatment and standard adjuvant chemotherapy in samples collected from Shanghai. The ratio of hypoxanthine to aspartate (Hyp/Asp) was observed to be significantly higher both in those samples collected from patients with outcomes of recurrence and/or death compared with patients free of recurrence or death ($p$ values of 0.003 and 0.033, respectively). With the Hyp/Asp ratios as the test variable and 5-year recurrence status as the state variable, a Receiver Operating Characteristic (ROC) curve was performed in the SPSS software (v20, IBM, Chicago, IL, USA). The AUC value was value of 0.722 (0.609-0.836, 95% confidence level) with a sensitivity of 0.875 and a specificity of 0.511 (Supporting figure 4A). The ROC curve was also performed with the ratio of Hyp/Asp and 5-year overall survival status as the test and state variables, which resulted in the AUC value of 0.694 (0.655-0.834, 95% confidence level) with a sensitivity of 0.563 and a specificity of 0.810 (Supporting figure 4B).

We further divided the samples equally into three groups based on their Hyp/Asp values (26 patients with Hyp/Asp values < 0.55, 26 patients Hyp/Asp values > 0.77, and 27 patients with values in between 0.55 and 0.77). There was a significantly longer time
to recurrence and overall survival between the first third of the patients (with Hyp/Asp values < 0.55) and the last third of the patients (with Hyp/Asp values > 0.77) on Kaplan-Meier analysis. Those patients with Hyp/Asp values < 0.55 had an average of time to recurrence of 57.9 months, while those with high Hyp/Asp values (> 0.77) had the time to recurrence of 36.6 months. The log rank $p$ value between these two groups was 0.002 (Supplementary Figure 4C). The overall survival time for those patients with low Hyp/Asp values was 62.6 months, and 51.8 months for patients with high Hyp/Asp values (log rank $p$ value = 0.024, Supplementary Figure 4D). All three groups (patients with low Hyp/Asp values, intermediate Hyp/Asp values, and high Hyp/Asp values) were further assessed with Kaplan-Meier analysis. The result showed significant difference in time to recurrence with $p$ value of 0.008, while the $p$ value for overall survival was non-significant ($p = 0.074$). The estimated time to recurrence and the overall survival time for those patients with intermediate Hyp/Asp values (between 0.55 and 0.77) was 44.7 and 58.6, respectively (Supplementary Figure 4E-4F).

4. Supporting figure legends

Supporting figure 1. A, PCA scores plot for CRC samples collected from Shanghai (13 components, R2Xcum=0.698); B, The result of a 999-time permutation test. All the R2 and Q2 in the permutation test are lower than the original ones and the intercept for the Q2 to the Y axe is below zero (Q2 intercept (0,-0.097)), indicating the validity of the OPLS-DA model; C, Representative chromatograms obtained from one CRC tissue (red line) and one adjacent non-tumor tissue (green line). The main figure was shown as total ion chromatogram (TIC). To enlarge the small peaks in the chromatogram, some metabolites (with star on the enlarged chromatograms) were shown as their molecular masses used in the current study. The masses for these metabolites were 220 for
cysteine; 174 for putrescine; 132 for myristate; 55 for palmitoleate; and 192 for kynurenine.

Supporting figure 2. PCA and OPLS-DA scores plot for CRC samples and adjacent non-tumor tissues collected from Hangzhou, Beijing and City of Hope. A, PCA scores plot for samples in the Hangzhou cohorts; B, PCA scores plot for samples in the Beijing cohorts; C, PCA scores plot for samples in the City of Hope cohorts; D, OPLS-DA scores plot for samples in the Hangzhou cohorts; E, OPLS-DA scores plot for samples in the Beijing cohorts; F, OPLS-DA scores plot for samples in the City of Hope cohorts.

Supporting figure 3. A, Fold change bar plot of differentially expressed genes between CRC tissues and adjacent normal ones. The fold changes were calculated with the mean values of CRC group to those of adjacent normal samples. * $p<0.05$, **$p<0.01$, ***$p<0.001$. $p$-values were calculated from Student’s t-test; B, Diagnostic results in four batches of CRC samples and gastric cardiac cancer. All of the probabilities were calculated with the same equation obtained from binary logical regression with CRC samples in the Shanghai cohort. Each red symbol represents one cancer tissue sample, while each blue symbol represents one adjacent non-tumor tissue sample.

Supporting figure 4. ROC and Kaplan-Meier curves for survival analysis in samples from Shanghai using the ratio of hypoxanthine to aspartate (Hyp/Asp).
References: