Supplementary method

Marker Selection

Selection of differentially abundant operational taxonomic units (OTUs)

We selected 1850 abundant OTUs present in at least 30% of all samples and calculated adjusted R^2 for each OTU. Before selecting the OTUs that are differentially abundant across stages of GC, we considered adjusting for potential confounding factors including age, gender, *Helicobacter pylori* status and tissue positions. The OTUs were divided into confounder-sensitive and -insensitive groups using linear regression with adjusted R^2.

First, we defined two variable sets:

Let $X = (X_1, X_2, \dots, X_p)^T$ *be the covariate matrix*, $F(x_k|c) = \Pr(X_k < x_k|c)$,

C: *the confounding factor set*

$A = \{X_k: F(x_k|c)$ *functionally depends on* c *for some* $c \in C,\ k = 1,2, \dots, p\}$,

$B = \{X_k: F(x_k|c)$ *does not functionally depends on* c *for some* $c \in C, k = 1,2, \dots, p\}$,

$X_k \in A$ is refered to be a confounder sensitive factor and $X_k \in B$ is refered to be a confouner insensitive factor.

Threshold Setting

The distribution of the adjusted R^2 is plotted in Graph a). In a further step, we set a threshold for the adjusted R^2 as the classification standard of set A and B. The threshold was obtained by adding some artificial auxiliary variables to the data as detailed in previously described.[1, 2] We randomly generated d auxiliary variables $\mathbf{z} \sim N_d(0, I_d)$ which were independently distributed. A relatively large d is preferred as a small d will lead to an unstable threshold. While d should not be too large since expectedly, a large d will introduce a large value to threshold R_k^2, which may make our thresholding method too strict to include some real

confounder sensitive factors. Considering these, a d that is relatively small and can also introduce low variability is preferred. We moved the number of auxiliary variables from 1 to 1000 and calculated the corresponding maximal adjusted R^2 (Graph b). Theoretically, the expectation of maximal adjusted R^2: $E\big(f(d)\big)$ is a monotone increasing function of the number of the auxiliary variables *d*. From graph b), the maximal adjusted R^2 tends to be stable when the number of auxiliary variables *d* is larger than 200. So, in a further step, we set *d* from 150 to 250 and for each *d*, we did 100 simulations and obtained the median, 25% and 75% quantile of $f(d)$ (Graph c). Graph c showed that the increment of $f(d)$ is very small as *d* increases when *d* is around 200. Thus, we used 200 auxiliary variables for thresholding the adjusted R^2. We simulated $f(200)1000$ times and used median value as our threshold value. Our covariate dataset extends to (X^T, Z^T) by adding the *d* auxiliary variables. Since **z** is truly confounder insensitive, we have $min_{X_k \in A} R_k^2 > max_{l=1,2,..,d} R_{p+l}^2$ and defined the confounder sensitive set *A* as below:

$$A = \{X_k : R_k^2 > max_{l=1,2,..,d} R_{p+l}^2, k = 1,2, \dots, p\}$$

Selection of OTUs that are sensitive to confounding factors

To evaluate the dependency of X_k on the confounding factors, we used multiple linear regression with the confounding factors mentioned above as the covariates and X_k as response. For each regression fitted with X_k as response, the corresponding adjusted R_k^2 was calculated. X_k with large adjusted R_k^2 was classified into *A*. With 200 auxiliary variables, 60 OTUs are classified as confounder sensitive and 1790 OTUs are classified as confounder insensitive. In a further step, we determined the significance of the confounder sensitive OTUs in $\hat{A}$.

Define $\hat{A} = (A_1, A_2, \dots, A_{p'})^T$. We determined the marginal significance of the covariates in $\hat{A}$ with logistic regression, where Y containing the cancer status served as the response and the A_k with the confounding factor matrix served as the predictors. Here, we inserted the confounding factors into our model to adjust

for their effects. After fitting the model, we obtained p value of each $X_k, k = 1,2,\dots,p'$, denoted as p_k and performed FDR adjustment of the values. Bacterial markers were finalized by selecting the OTUs with adjusted p_k smaller than 0.05.

Selection of OTUs that are insensitive to confounding factors

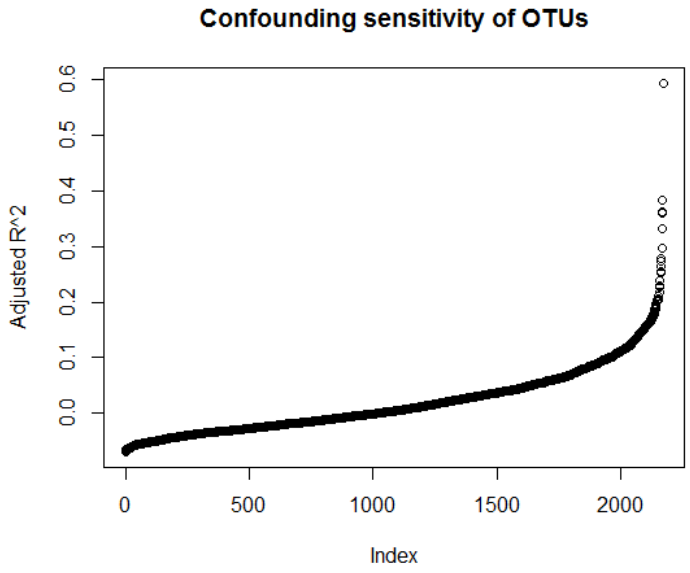
For the confounding-insensitive covariates, we considered using some screening methods first due to the high dimensionality of the confounder insensitive covariates . We adopted the Model-Free Feature Screening earlier proposed[3] to screen out a candidate set of the OTU markers. Define $\omega_k = E\{\Omega_k^2(Y)\}, where\ \Omega(y) = E[\mathbf{x}(y|\mathbf{x})], k = 1,2,\dots,p; Y: the\ response\ vector$. We used the cancer status as the response. ω_k served as the marginal utility measure for ranking the covariates.

Before the screening, all the sample covariates were standardized to $\frac{1}{n}\sum_{i=1}^{n} X_{ik} = 0$ and $\frac{1}{n}\sum_{i=1}^{n} X_{ik}^2 = 1$ for $k = 1,2,\dots,p$. The natural estimator for ω_k is $\widehat{\omega_k} = \frac{1}{n}\sum_{i=1}^{n}\{\frac{1}{n}\sum_{j=1}^{n} X_{ik}\mathbf{1}(Y_i < Y_j)\}^2$. $\widehat{\omega}$ was used in our case for feature screening. A threshold was also needed to determine the candidate set of bacterial markers. A hard cutoff was chosen by retaining a fixed number of covariates from screening. According to the thresholding rule proposed by Fan and Lv (2008),[4] we set the number at $[n/\log n]$. Similar to the selection of the confounder sensitive variables, the logistic regression was used to determine the significance of the covariates. We also performed p value adjustment by FDR and chose the covariates (OTUs) with adjusted p value smaller than 0.05.

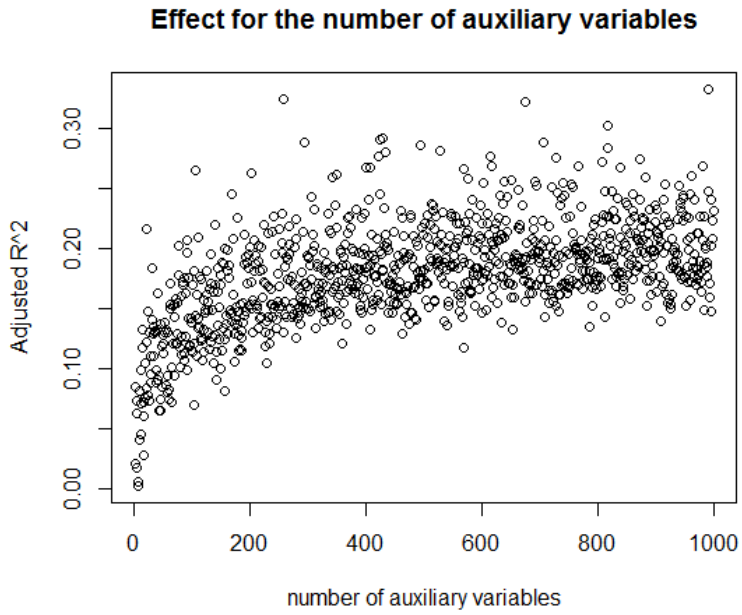
Abundance Adjustment

Since some OTU abundance is closely correlated with the confounding factors included in our study, we considered adjusting the OTU's abundance with respect to confounding factors. To keep the consistency of our analysis, linear regression was used to adjust the confounding effect with the OTU abundance after log transformation as the response and the confounding factors as the predictors. The residuals of the

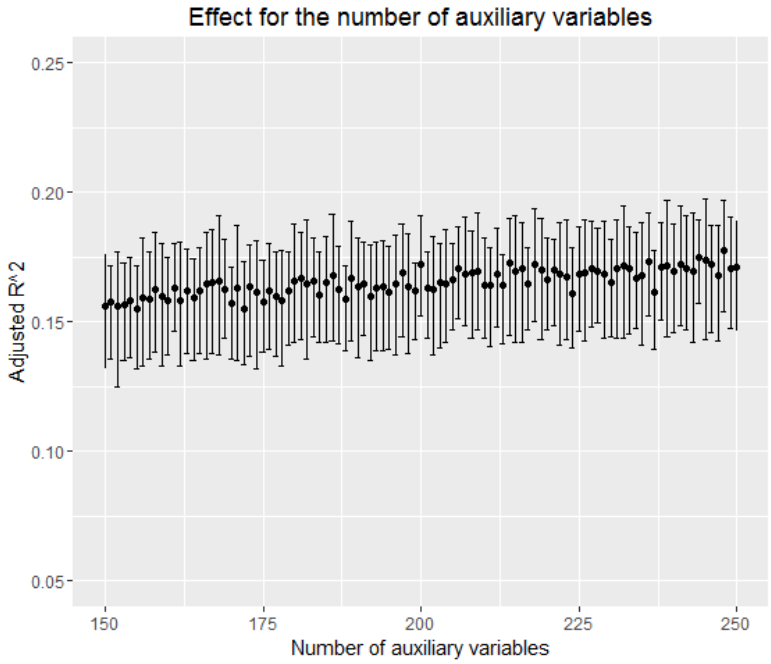
regression are linearly independent of the confounding factors showing abundance changes after removing the effect of the confounding factors.



Graph a)



Graph b)



Graph c)

REFERENCES

1. Luo XH, Stefanski LA, Boos DD. Tuning variable selection procedures by adding noise. Technometrics 2006;48:165-175.
2. Wu YJ, Boos DD, Stefanski LA. Controlling variable selection by the addition of pseudovariables. Journal of the American Statistical Association 2007;102:235-243.
3. Zhu L, Li L, Li R, et al. Model-Free Feature Screening for Ultrahigh Dimensional Data. J Am Stat Assoc 2011;106:1464-1475.
4. Jianqing Fan, Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. 2008 Royal Statistical Society 1369–7412/08/70849J. R. Statist. Soc. B (2008)70,Part 5, pp. 849–911