# Methods

## Datasets and networks

Multiplatform genomics datasets included gene expression profiles and somatic mutation in MAF (Mutation Annotation Format) files were downloaded from The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov/). And the clinical data were obtained through the TCGA Data Commons (https://gdc.cancer.gov/). Two other validation datasets were downloaded from Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/gds) with accession number GSE21510 and GSE39582. Besides, it was used as a background network that the functional association network with high confidence (experiment lab score>300) which includes direct (physical) and indirect (functional) associations obtaining from the STRING database (version 10.5, <http://string-db.org>) [[15](#_ENREF_15)].

## Visualizing and summary of mutation datasets

The mutations from MAF files were visualized and summarized through summary plots and oncoplots using the R/Bioconductor maftools package [[16](#_ENREF_16)].

## Constructing SSN (Sample-Specific Network) for each sample

An SSN for each sample was constructed by a sample-specific network (SSN) method [[14](#_ENREF_14)], which can infer individual-specific networks based on the expression data of a single sample from the following strategies. Firstly, the normal samples were considered as reference samples and a reference network was obtained by computing Pearson correlation coefficient (*PCC*) of each pair of molecules as an edge in the background network, which was conducted from STRING protein-protein interaction (PPI) network. And then, a perturbed network was constructed by adding a single sample to the reference samples and computing *PCC*s again. Finally, edges were kept to construct an SSN for this single sample if they showed statistically significant differential *PCC*s () based on the evaluation of SSN theory when comparing the perturbed network with the reference network.

## Functional network identification for cancer

Specific SSN for each tumor sample was obtained by deleting the edges presented in normal samples. If an edge appeared in more than 90% SSNs of tumor samples, the edge would be collected to form a functional network for COAD. The nodes and edges in the functional network were used as representative features for COAD, which were named as functional genes and functional interactions of COAD, respectively.

## The Enrichment Analysis of GO and KEGG Pathway

The enrichment analysis of GO and KEGG pathway for functional genes were performed using DAVID web service (<https://david.ncifcrf.gov/>) [[17](#_ENREF_17), [18](#_ENREF_18)] with specifying a *p*-value<0.05 for statistical significance.

Furthermore, genes in five known cancer gene sets were used as a proxy for the potential cancer-related genes including the curated gene sets in pathway in cancer (hsa05200), colorectal cancer (hsa05210), cancer gene census [[19](#_ENREF_19)], pan-caner driver genes [[20](#_ENREF_20)], and cancer driver genes [[21](#_ENREF_21)]. And the probability *p*-values that can reflect whether functional genes are significantly enriched in these known cancer gene sets were calculated by the following formula [[22](#_ENREF_22)]:

 *p*-value=, (1)

where  is the total number of genes of the human genome,  is the number of genes in a known cancer gene set,  is the number of functional genes,  is the number of overlapping genes between functional genes and the known cancer gene set. If the *p*-value is less than 0.05, then it means that the functional genes are significantly enriched in the known cancer gene set. And then Venn diagrams were used to display the relationship between functional genes and the five known cancer gene sets.

## Functional genes as diagnostic biomarkers

To check whether functional genes can be used as diagnostic biomarkers for colon adenocarcinoma, 5-fold cross-validation was conducted to perform normal/tumor classification by a support vector machine (SVM), which was implemented in R with function ‘ksvm’ in ‘kernlab’ package. And the receiver operating characteristic (ROC) curve was drawn by R using the ‘ROCR’ package. In detail, TCGA data were used as training and test set, and GSE21510 data were used as an independent external validation dataset. To settle the problem of data imbalance, TCGA tumor data were divided into subgroups to make sure each subgroup had almost the same sample size with TCGA normal dataset. And then SVM model with 5-fold cross-validation was performed for each tumor subgroup and normal samples. Furthermore, hierarchical clustering was performed by using the gene expression of functional genes in both tumor and normal samples. And then heat maps were used to show the results.

## Colon adenocarcinoma subtypes and survival analysis

Colon adenocarcinoma samples were divided into subtypes by consensus clustering algorithm [[23](#_ENREF_23)] using the expression data of functional genes. Consensus clustering was performed by ConsensusClusterPlus R-package using 1000 iterations, 80% sample resampling from 2 to 10 clusters, Ward linkage and the distance of Pearson correlation coefficient. Then one clustering solution was selected as a subtype solution. Differentially expressed genes (DEGs) associated with each subtype were identified by carrying out a two-sided *t*-test for each gene by comparing this subtype with the rest subtypes, and then the unique top 100 upregulated DEGs and downregulated DEGs with the lowest *p*-value were selected as representative genes for each subtype. Then their enriched biological processes and KEGG pathways were compared using the R package ‘clusterProfiler’ [[24](#_ENREF_24)] which can compare biological themes among gene clusters. Subtype-specific clinical features and somatic alteration features were also assessed. Besides, Kaplan-Meier survival curves were drawn for subtypes and log-rank *p*-values were computed using the R package ‘survival’ [[25](#_ENREF_25)].

## Prognostic prediction of COAD using functional genes and interactions

Association of functional genes and interactions with patients’ overall survival were assessed by Kaplan-Meier survival curves and the log-rank tests. Based on the expression level of functional genes orof interactions, samples were divided into two subgroups with low- and high- expression. And then a univariate Cox regression analysis was done for each functional gene and interaction. Furthermore, a multivariate Cox regression analysis was further done to investigate and control the influences of the confounders on functional genes or edges with *p*-value less than 0.05 in the univariate Cox regression analysis. The confounders included sex information, pathologic stages, retrospective collection indicator, race, the year of initial pathologic diagnosis, age at initial pathologic diagnosis and microsatellite status. Functional genes and interactions were identified as prognosis biomarkers for cancer when they showed significant differences between the low- and high- expression subgroups in both univariate and multivariate cox analysis. The function ‘coxph’ in R was used to do this job.