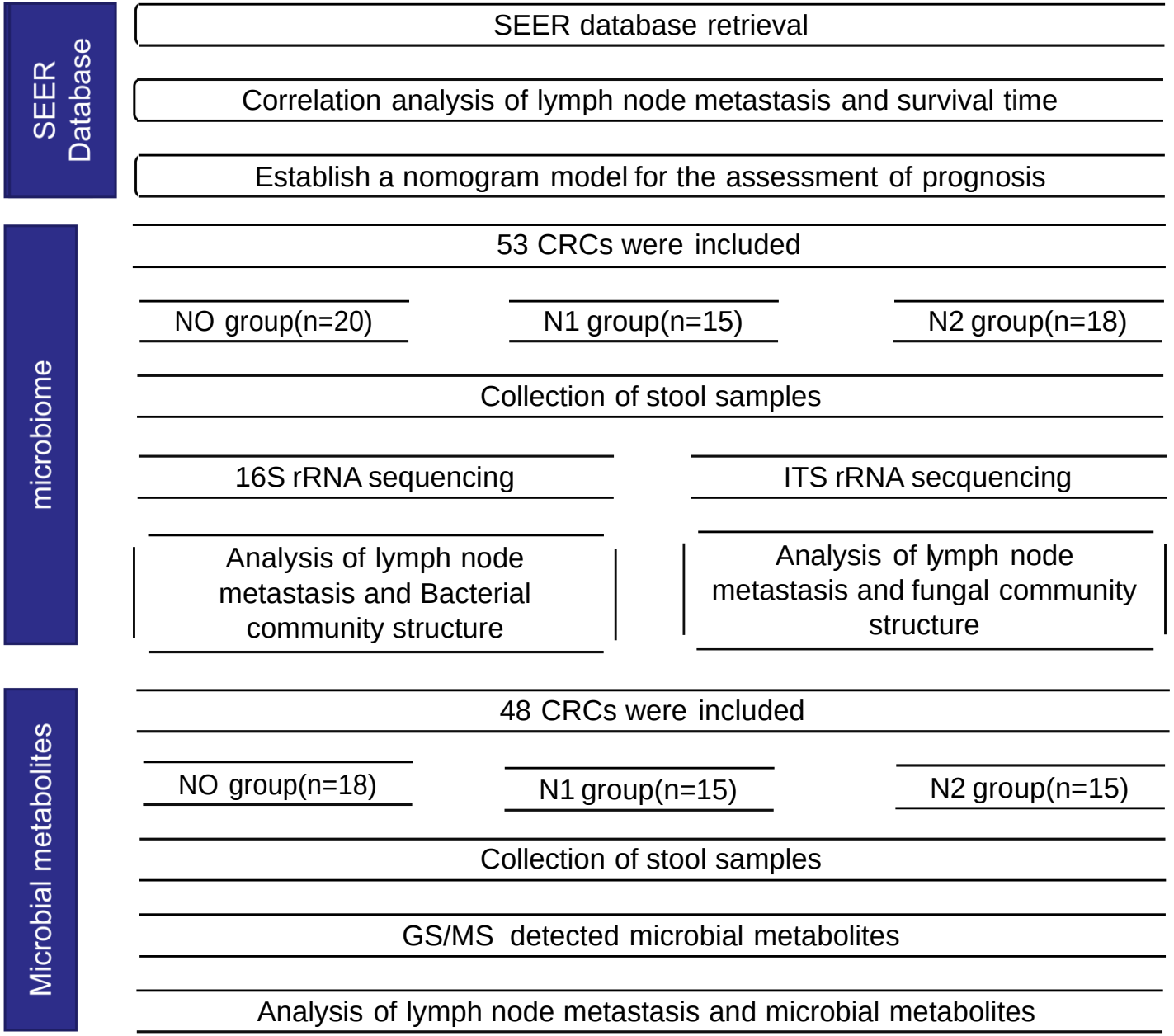
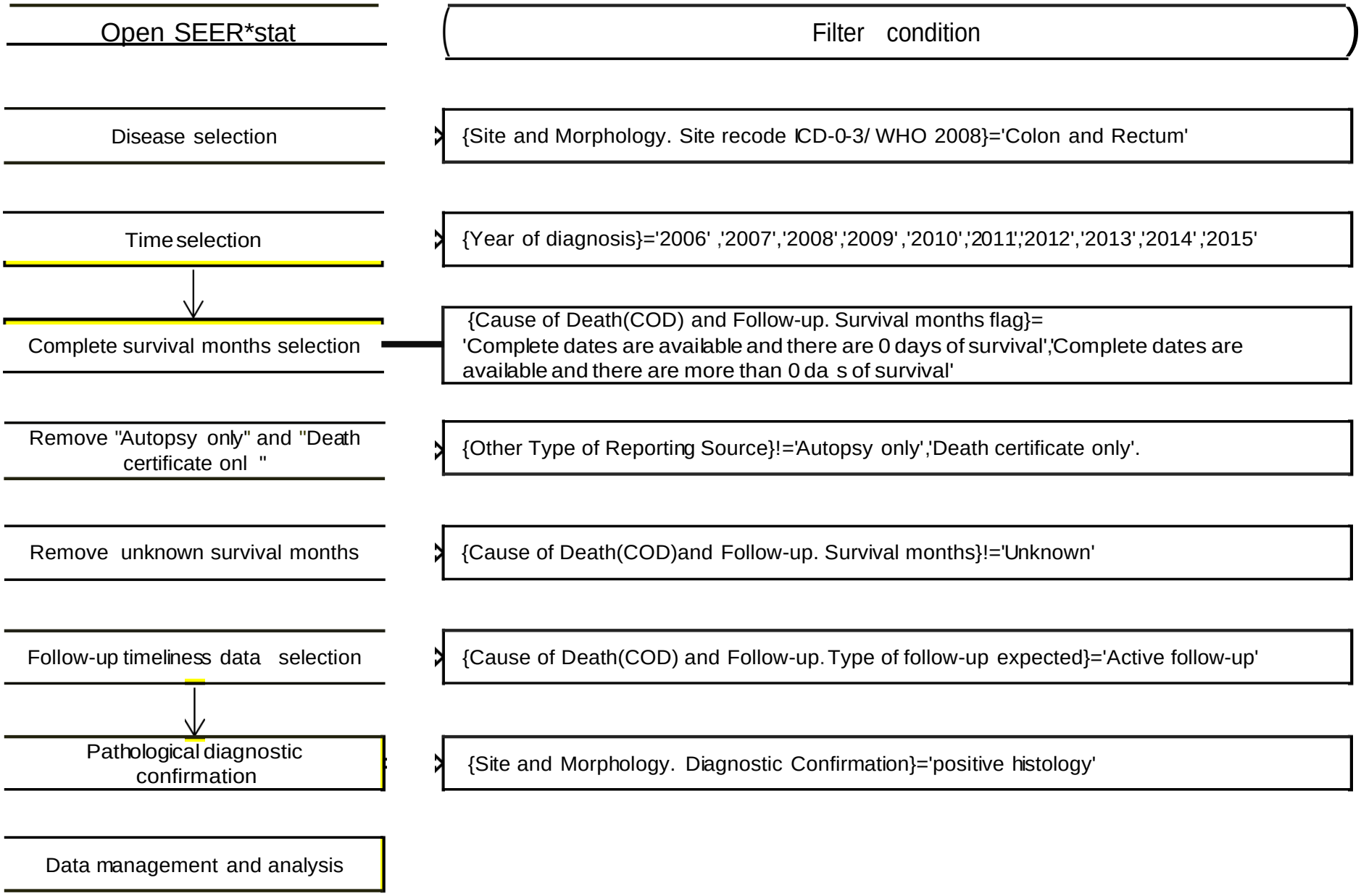
Supplemental material Figure S1. Study strategy

The SEER database was used to analyze the effect of lymph node metastasis status on prognosis of CRCs. After screening and identification, 53 stool samples from CRC patients with different lymph node metastasis status were used to detect gut microbiome and 48 stool samples used recruited to detect gut microbial metabolites. Bioinformatics analysis was executed to analyze the community structure and microbial metabolites differences associated with lymph node stages.

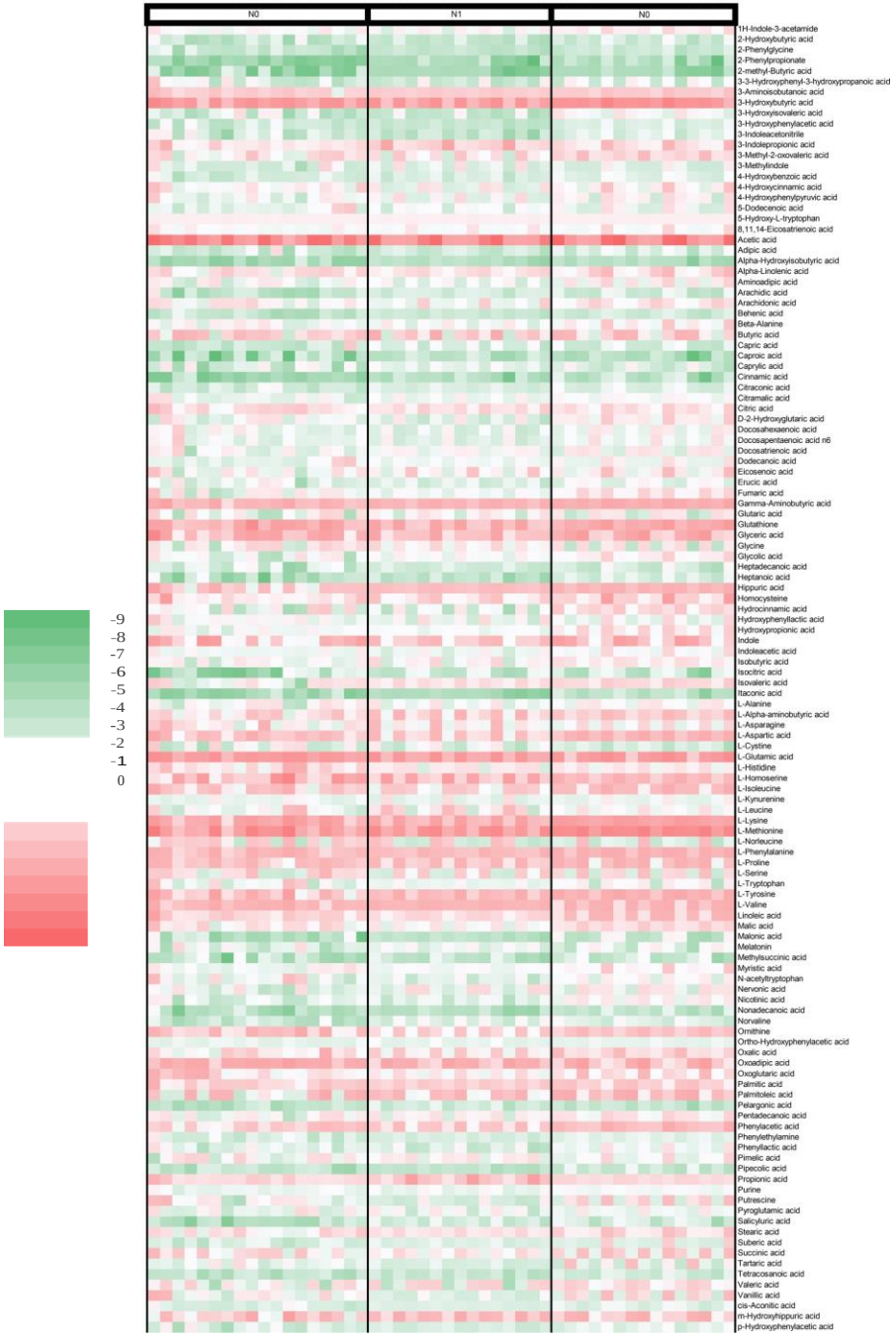


Supplemental material Figure S2. SEER database retrieval strategy

SEER*Stat software (seer.cancer.gov/seerstat) was used to access the SEER data after signing a research data agreement. All patients diagnosed with colon and rectal adenocarcinoma from 2006 to 2015 were included. The relationship between lymph node metastasis status and prognosis was analyzed according to the retrieved data.



Supplemental material Figure S3 The heatmap shows the concentrations of all 124 metabolites of stool samples from the CRCs with different lymph node metastasis stages.



Supplemental material 1. Gut microorganism detection methods

1. DNA extraction and PCR amplification

An E.Z.N.A.® Stool DNA Kit (Omega Bio-Tek, Norcross, GA, U.S.) was used to extract total DNA from the stool samples. The quality of the purified DNA was determined by the nanodrop ND-1000 spectrophotometer (LabTech, Washington, DC,USA) with absorbances at 260 nm and 280 nm (A260/A280). Electrophoresis with a 2.0% (w/v) agarose gel was used to verify the integrity of DNA. The V3–V4 region of the bacterial 16S ribosomal RNA gene (the primers of 16S V3–V4 rDNA were as follows: forward, CCTACGGGNGGCWGCAG; reverse, GACTACHVGGGTATCTAATCC) and fungal ITS ribosomal RNA gene (using the primers 1F 5'-CTTGGTCATTTAGAGGAAGTAA-3' and 2R 5'-GCTGCGTTCTTCATCGATGC-3') were amplified by PCR. PCR amplification was performed at 95℃ for 3 min, 25 cycles at 95℃ for 30 s, 55℃ for 30 s, and 72℃ for 45 s and a final extension at 72℃ for 5 min. PCR amplifications used a 25 μl volume containing 5 μl of DNA template, 2 μl of Nextera XT Index Primer 1 (10 M), 2 μl of Nextera XT Index Primer 2 (10 M), and 16 μl of ddH_2O. PCR for each gene used three biological replicates, with three technical replicates per experiment. The amplicons were extracted from 2% agarose gels, purified using the AxyPrep DNA Gel Extraction Kit (Axygen Biosciences, Union City, CA, USA) and quantified using QuantiFluor™-ST.

2. MiSeq Library construction and sequencing

Qubit® 3.0 (Life Technologies, Invitrogen) was used to quantify the purified PCR products. The products were ligated with Y adapter and the self-ligated Y adapters were taken out by using magnetic nanoparticles. The pooled DNA products constructed an Illumina Pair-End library and the amplicon library was pair-end sequenced (2×250) on an Illumina MiSeq platform (Shanghai BIOZERON Co., Ltd.) according to standard protocols. The raw reads were deposited into the NCBI Sequence Read Archive (SRA) database (Huzhou Central Hospital : SUB4802613 SRP151510 :PRJNA478277; Huzhou Central Hospital : SUB4648640 SRP169843 : PRJNA506089).

3. Sequencing data bioinformatics analysis

The Silva database (Release119 http://www.arb-silva.de)[1] and the Unite database (Release 6.0 http://unite.ut.ee/index.php)[2] were used for comparison of 16S rRNA gene sequences and ITS gene sequences, respectively. Using Mothur software[3] (http://www.mothur.org/wiki/Main_Page) according to the criteria at http://en.wikipedia.org/wiki/Fastq, sequencing data was processed and optimized. The specific operation was as follows: (i) cutadapt (version 1.11) was used to remove sequences without primers, and the primers matching process allowed an error of 0.15, (ii) pandaseq (version 2.9) was used to assemble PE reads and required an overlap longer than 10 bp, (iii) reads receiving an average quality score <20 were discarded, (iv) Mosaic sequences and sequences outside 300–480 bp were discarded.

The number of species in each sample was estimated by OTU (Operational Taxonomic Units) with 97% similarity. RDP-classifier software annotates the sequences for species and OTU tables were generated into species abundance tables at

different classification levels (phylum, class, order, family, genus) by using Qiime software. The complete linkage hierarchical clustering technique with the R package HCLUST (http://sekhon.berkeley.edu/stats/html/hclust.html) was used to perform the cluster analysis.

References

1. Quast, C., et al., *The SILVA ribosomal RNA gene database project: improved data processing and web-based tools.* Nucleic Acids Res, 2013. **41**(Database issue): p. D590-6.
2. Koljalg, U., et al., *Towards a unified paradigm for sequence-based identification of fungi*. Mol Ecol, 2013. **22**(21): p. 5271-7.
3. Schloss, P.D., D. Gevers, and S.L. Westcott, *Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies.* PLoS One, 2011. **6**(12): p. e27310.